AI STACK – ข้อกำหนดทางเทคนิค

เอกสารฉบับนี้ระบุข้อกำหนดด้านการทำงานและข้อกำหนดทางเทคนิคขั้นต่ำสำหรับการติดตั้งและใช้งาน AI STACK แพลตฟอร์มการจัดการทรัพยากร (Resource Management Platform) เพื่อบริหารจัดการทรัพยากร GPU/CPU/Memory, งานแบบ Containerized Workloads และบริการคอมพิวติ้งด้าน AI ในสภาพแวดล้อมที่มีความเสถียรสูง และสามารถทำงานได้ตลอดเวลาโดยที่ไม่เกิดการล่มหรือ Downtime. (High Availability)

1. การขออนุมัติและการสมัครโครงการ (Project Application and Approval)

- แพลตฟอร์มต้องสามารถรองรับการสมัคร ขออนุมัติ การเปิดใช้งาน และการยกเลิกโครงการ (Project) ได้
- เมื่อโครงการได้รับอนุมัติแล้ว ต้องสามารถกำหนดค่าพารามิเตอร์การใช้งาน GPU ได้
- การสมัครที่อยู่ในขอบเขตการกำหนด สามารถดำเนินการต่อได้โดยไม่ต้องผ่านการอนุมัติจากผู้ดูแลระบบ (Admin)

2. การจัดการโครงการและผู้ใช้งาน (Project and User Management)

- ผู้ดูแลระบบสามารถสร้าง แก้ไข และลบเนื้อหาโครงการและสมาชิกได้
- ต้องมีฟังก์ชันการลบโครงการแบบกลุ่ม (Batch Deletion)
- ผู้ดูแลระบบสามารถกำหนดข้อจำกัดการใช้ GPU บางรุ่นสำหรับโครงการ และผู้ใช้โครงการจะเห็นเฉพาะ GPU รุ่นที่ได้รับอนุญาตเท่านั้น
- สามารถรองรับผู้ดูแลระบบหลายระดับ
- สามารถการบันทึกประวัติการใช้งานของผู้ใช้และผู้ดูแลระบบในทุกการดำเนินการ
- ผู้ดูแลระบบสามารถปรับการจัดสรรทรัพยากร (GPU/CPU/RAM Quota) สำหรับแต่ละโครงการได้

3. การจัดการ Image และการปรับใช้ (Image and Deployment Management)

- รองรับการดึงอัตโนมัติ (Automatic Image Pulling) จาก NVIDIA GPU Cloud (NGC) เพื่อให้ผู้ดูแลสามารถสร้างแม่แบบบริการ (Service Templates) ได้
- รองรับการนำเข้า Image จาก NGC และ Docker Hub
- ผู้ดูแลระบบสามารถเลือกการเข้าถึง Container ผ่าน SSH ทั้งแบบรหัสผ่านและแบบ Key-based Authentication ได้
- รองรับการเชื่อมต่อบัญชีผู้ใช้กับ AD, LDAP หรือ SSO
- ต้องมี Dashboard แบบกราฟิก แสดงการจัดสรรทรัพยากร GPU ใน 2 โหมด: Aggregation Mode และ Shared Mode

4. การแบ่งพาร์ติชัน GPU และการจัดการต้นทุน (GPU Resource Partitioning and Cost Management)

- รองรับความสามารถ GPU Slicing เพื่อแบ่งหน่วยความจำของ GPU แต่ละรุ่น และกำหนดทรัพยากร CPU/Memory ให้ผู้ใช้เลือกได้
- รองรับการคำนวณค่าใช้ GPU รายชั่วโมง
- ผู้ดูแลระบบสามารถกำหนดโควตาการใช้งานรายโครงการได้ โดยเลือกเพิ่มโควตาหรือกำหนดให้ใช้งานไม่จำกัดได้
- เมื่อโครงการใดใช้เกินโควตาที่กำหนด ระบบต้องแจ้งเตือนในขั้นตอนการสร้าง Container

5. การตรวจสอบและการรายงาน (Monitoring and Reporting)

- ต้องมี Dashboard แบบรวมศูนย์ (Centralized Graphical Dashboard) เพื่อตรวจสอบการใช้งานฮาร์ดแวร์ของ GPU Servers (CPU, RAM, GPU, GPU RAM)

- รองรับการจัดตารางการใช้งานทรัพยากรแบบ Reservation-Based Scheduling เพื่อให้ผู้ใช้สามารถจอง GPU/CPU/DRAM ตามช่วงเวลาได้

- หากไม่มีทรัพยากรเพียงพอ ระบบต้องแจ้งผู้ใช้และแสดงรายการรอ (Waiting Queue) ให้ผู้ดูแลระบบเห็น

6. การสลับเวอร์ชัน CUDA (CUDA Version Switching)

- ต้องมีอินเทอร์เฟซแบบกราฟิก(GUI)สำหรับผู้ดูแลในการกำหนดและสลับเวอร์ชัน CUDA ในแต่ละ Node ได้

7. การจัดเก็บและความปลอดภัย (Storage and Security)

- รองรับการเชื่อมต่อกับ MinIO Object Storage

- ต้องมีการสแกน Container เพื่อหาช่องโหว่ด้านความปลอดภัยตามรอบเวลา

- รองรับการเชื่อมต่อกับ Label Studio สำหรับการทำ Data Labeling

- ต้องรองรับการจัดการทรัพยากร GPU ทั้ง NVIDIA และ AMD

8. ความพร้อมใช้งานสูง (High Availability – HA) และ RCS (Rapid Container Service)

- รองรับการออกแบบระบบ HA เพื่อลดความเสี่ยงจากการล้มเหลวของจุดใดจุดหนึ่ง (Single Point of Failure)

- มี RCS (Rapid Container Service) ทำให้ผู้ใช้สามารถปรับใช้งานโมเดล/อิมเมจได้โดยไม่ต้องใช้ GPU สำหรับงาน Inference

9. สภาพแวดล้อมการพัฒนา AI ที่ติดตั้งล่วงหน้า (Pre-installed AI Development Environment)

- แพลตฟอร์มมาพร้อมสภาพแวดล้อมการพัฒนา Deep Learning ที่สมบูรณ์ ได้แก่: Ubuntu 22.04 LTS หรือใหม่กว่า, NVIDIA Driver, NVIDIA CUDA 11.5 หรือใหม่กว่า, cuDNN, NVIDIA Docker, Jupyter

- และในระยะเวลารับประกัน สามารถได้รับการอัปเดตอย่างต่อเนื่องสำหรับ OS และซอฟต์แวร์ดังกล่าว

10. การฝึกอบรมแบบยืดหยุ่น (Elastic Training)

- รองรับ Elastic Training ที่สามารถปรับขยายหรือลดจำนวน Worker Containers ระหว่างการฝึกอบรมได้ตามทรัพยากรใน Cluster

- เพื่อเพิ่มประสิทธิภาพการใช้ทรัพยากรให้เหมาะสมที่สุด

11. การอนุญาตสิทธิการใช้งาน (Licensing)

- ซอฟต์แวร์เชิงพาณิชย์ของแพลตฟอร์มต้องให้ภายใต้สิทธิการใช้งานถาวร (Perpetual License – Buy-out Model)

- สิทธิการใช้งานต้องครอบคลุมการเข้าถึงแพลตฟอร์มอย่างต่อเนื่องภายหลังครบกำหนดระยะเวลา 3 ปีแรก