
Nontri AI Gateway
การใช้ทรัพยากร HPC ของมหาวิทยาลัยเกษตรศาสตร์ สำหรับงาน AI จริง

NotebookLM

ทำไมต้อง HPC สำหรับ AI?

Constraint

- AIใช้พลังคำนวณสูง
- Deep Learning = คำนวณมหาศาล
- เครื่องส่วนตัวไม่พอ

Possibility

NotebookLM

HPC คืออะไร

คลัสเตอร์คอมพิวเตอร์จำนวนมาก

มีGPUหลายตัว

บริหารจัดการด้วยระบบ Queue

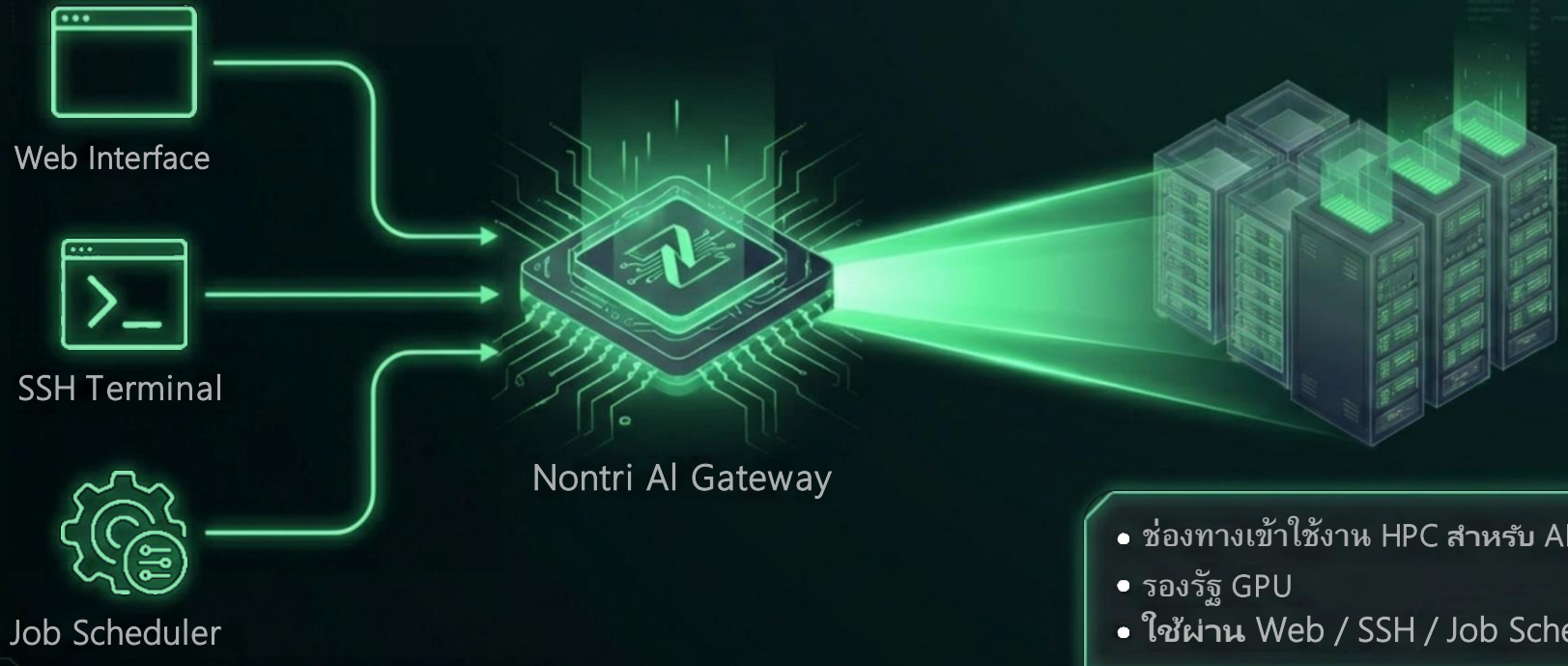
NotebookLM

HPC ในประเทศไทย

RESOURCE	NONTRI AI (มก.)	ERAWAN (มช.)	LANTA (สวทช.)
CPU	496 Cores	384 Cores	20,480 Cores
RAM	11 TB	6 TB	10 x 4 TB
GPU	A100 (80G) - 32 GPU	A100 (80G) - 24GPU H100 - 4GPU	A100 - 704 GPU
NETWORK	2x200 Gb InfiniBand	Infiniband HDR 200 Gbps, Latency 90 ns, Throughput 16 Tbps	-
STORAGE	SSD 90 TB HDD 670 TB	NVME 150 TB SAS 588 TB	945TB NVMe HDD 10 PB
Price	Free	ภายใน CPU 0.25 บาท/ชม. GPU 5 บาท/ชม. ภายนอก CPU 5 บาท/ชม. GPU 200 บาท/ชม.	บริการ compute:hr / GPU:hr -ภาครัฐ 15/45 -วิสาหกิจ 45/135 -ระหว่างประเทศ 70/210

NotebookLM

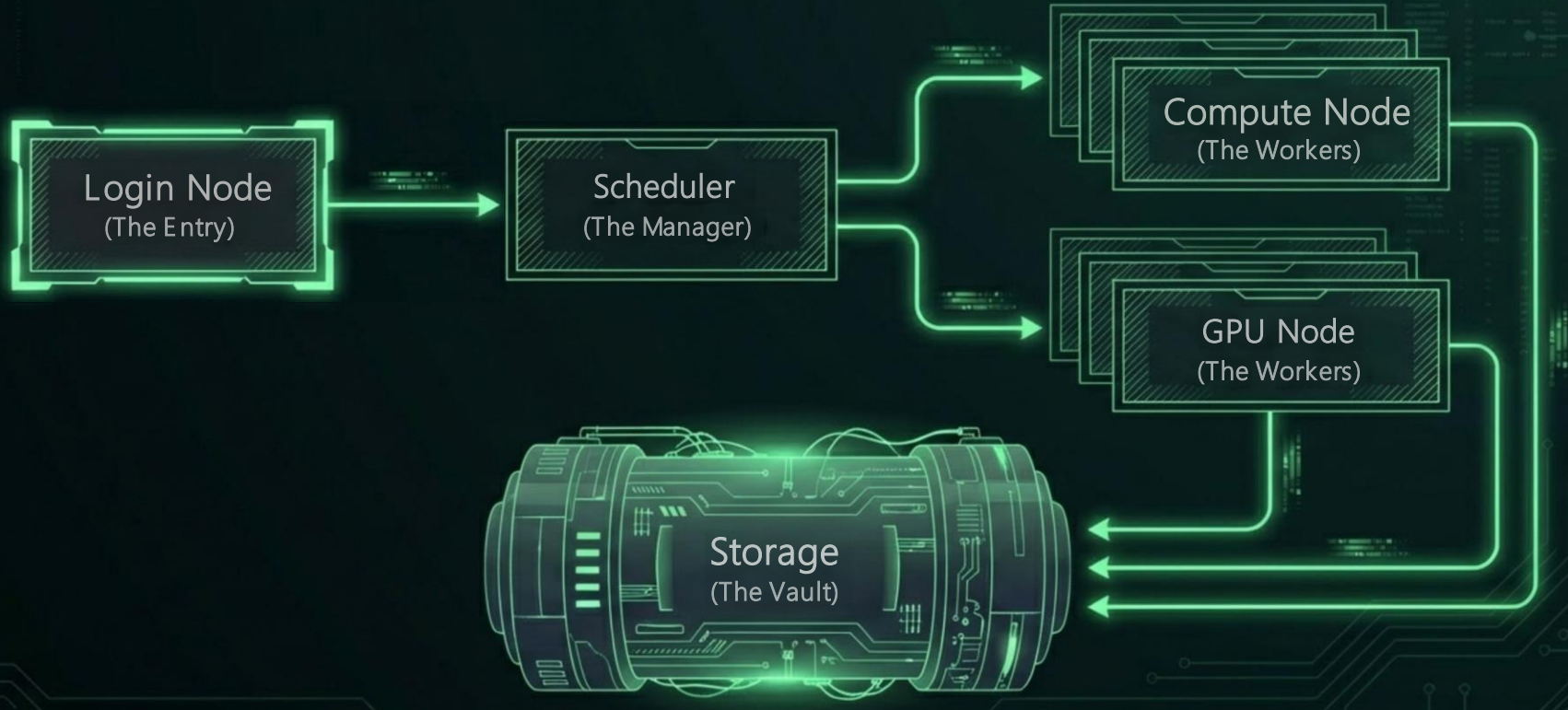
Nontri AI Gateway คืออะไร?



- ช่องทางเข้าใช้งาน HPC สำหรับ AI
- รองรัฐ GPU
- ใช้ผ่าน Web / SSH / Job Scheduler

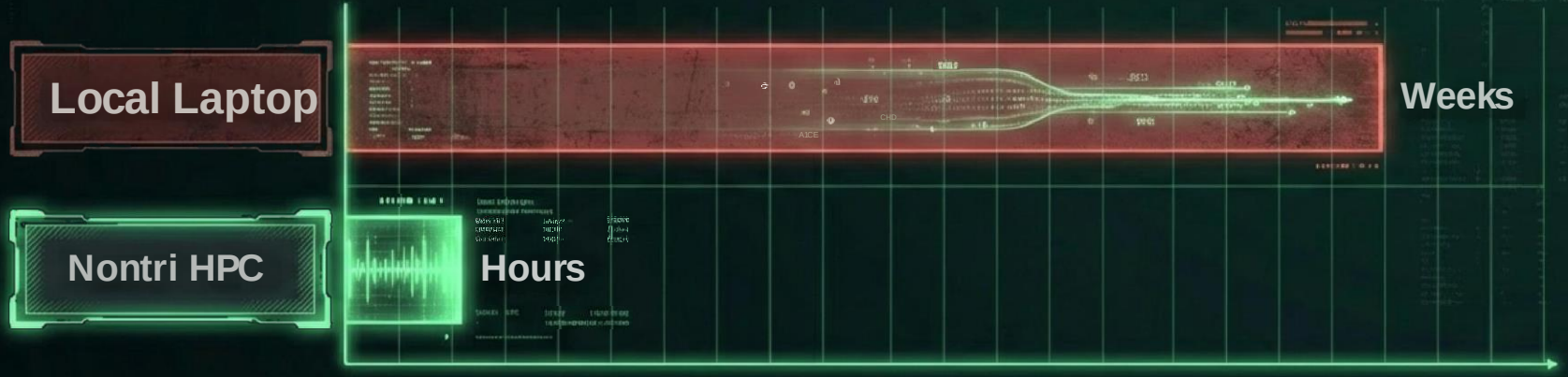
NotebookLM

โครงสร้างเบื้องต้นของระบบ



NotebookLM

ทำไมควรใช้ ?



- ไม่ต้องซื้อเครื่องแรง
- เทรนโมเดลใหญ่ได้
- ทำโปรเจกต์ระดับแข่งขันได้จริง

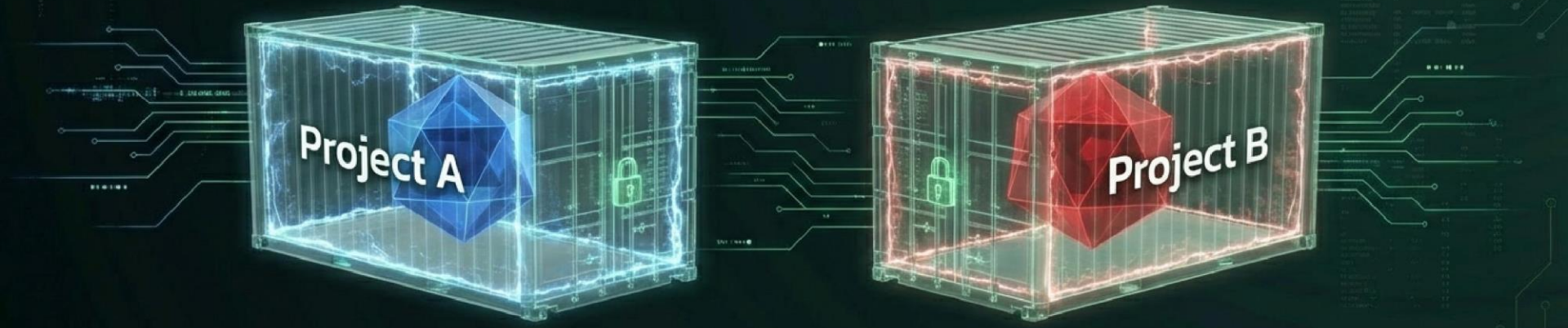
NotebookLM

ขั้นตอนที่1: การเข้าใช้งาน

- สมคร Account: https://nontriai.ku.ac.th
- Login ผ่าน Web/SSH

แจกฟรี 100 Token สำหรับการ Run งาน

NotebookLM



ขั้นตอนที่ 2: การจัดการ Environment

- ใช้ Conda / venv
- แยก environment ต่อโปรเจกต์
- ติดตั้ง PyTorch / TensorFlow

NotebookLM

ขั้นตอนที่ 3: การส่ง Job

- ห้ามรันงานหนักบน Login Node
- ใช้ Job Scheduler
- ขอ GPU ตามจำนวนที่ต้องการ

NotebookLM

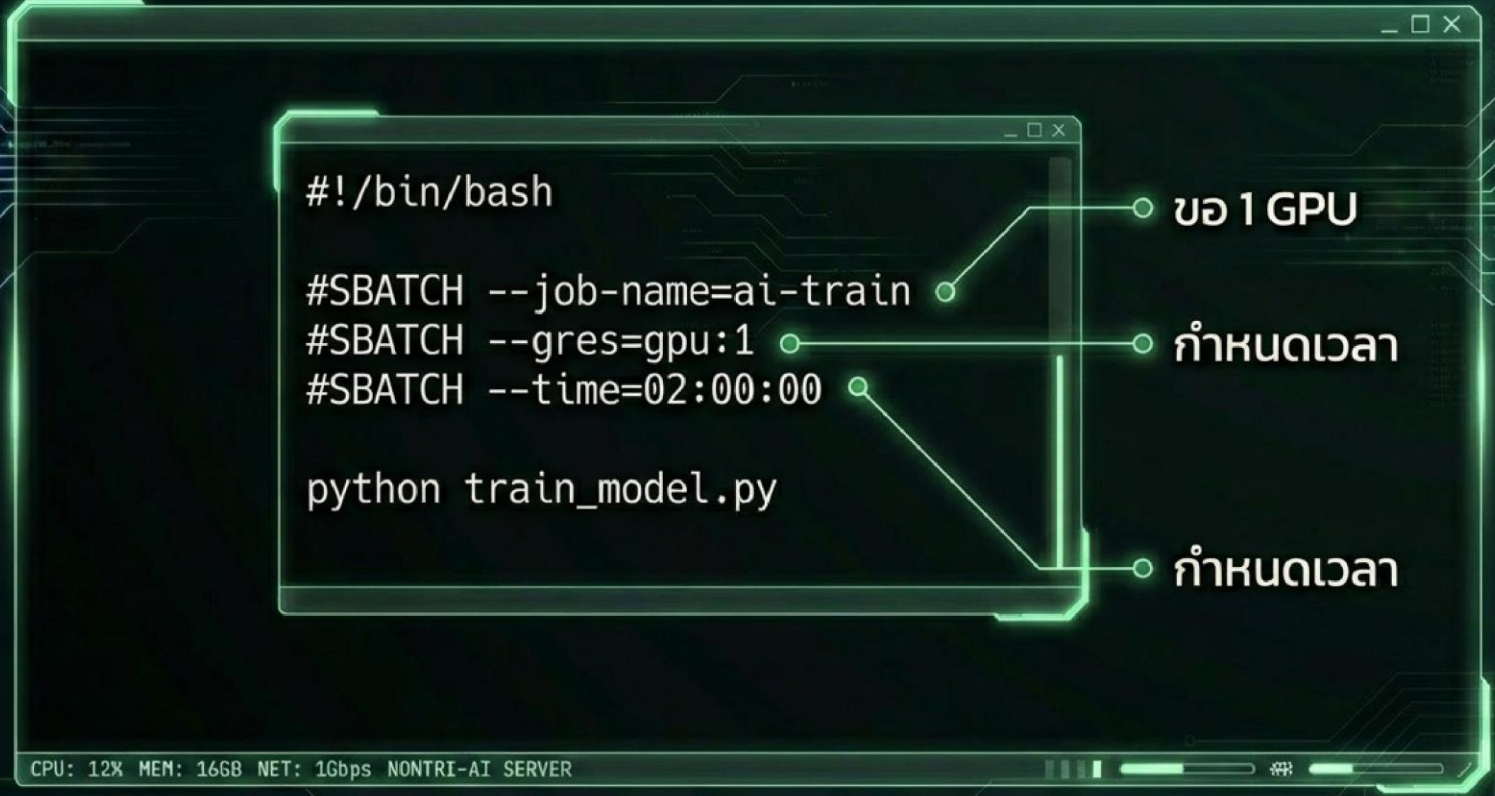
ตัวอย่างคำสั่ง Environment

```
user@nontri-ai:~$ conda create -n my_env python=3.9
Collecting package metadata (current_repodata.json): done
Solving environment: done
# ...
To activate this environment, use
    $ conda activate my_env
To deactivate an active environment, use
    $ conda deactivate

user@nontri-ai:~$ conda activate my_env
(my_env) user@nontri-ai:~$ pip install torch torchvision
Collecting torch
  Downloading torch...
Collecting torchvision
  Downloading torchvision...
Installing collected packages: torch, torchvision
Successfully installed torch torchvision
(my_env) user@nontri-ai:~$ _
```

CPU: 12% MEM: 16GB NET: 1Gbps NONTRI-AI SERVER

NotebookLM

ตัวอย่างการ submit job (sbatch)



NotebookLM

GrpTRES GrpTRESMins

- ตรวจสอบสถานะ (Status)
- ดู Log files
- ตรวจ GPU usage

```
      GrpTRES =
            cpu:        Limit = 60, current value = 0
            mem:        Limit = 131072, current value = 0
       gres/gpu:1g.10gb:        Limit = 1, current value = 0
       gres/gpu:2g.20gb:        Limit = 1, current value = 0
       gres/gpu:3g.40gb:        Limit = 1, current value = 0
       gres/gpu:7g.80gb:        Limit = 1, current value = 0

GrpTRESMins =
            cpu:        Limit = 0, current value = 0
       gres/gpu:1g.10gb:        Limit = 0, current value = 0
       gres/gpu:2g.20gb:        Limit = 0, current value = 0
       gres/gpu:3g.40gb:        Limit = 0, current value = 0
       gres/gpu:7g.80gb:        Limit = 0, current value = 0
```

Log Output

```
      GrpTRES =
            cpu:        Limit = 60, current value = 0
            mem:        Limit = 131072, current value = 0
       gres/gpu:        Limit = 1, current value = 0

 GrpTRESMins =
            cpu:        Limit = 0, current value = 0
       gres/gpu:        Limit = 0, current value = 0
```

CPU: 18% MEM: 24GB NET: 2Gbps NONTRI-AI SERVER

NotebookLM

Partition

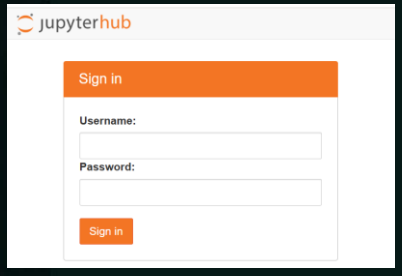
- ตรวจสอบสถานะ (Status)
- ดู Log files
- ตรวจ GPU usage

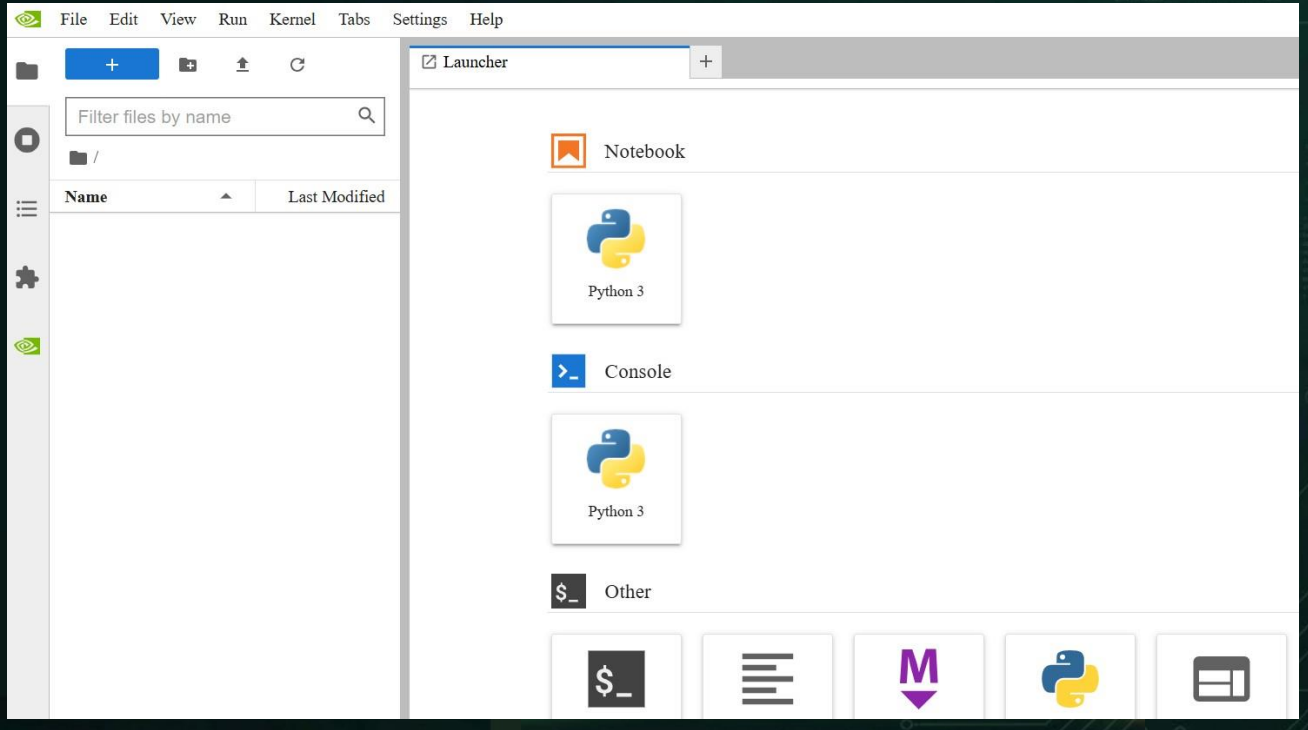
```
PartitionName=fullq
   AllowGroups=ALL AllowAccounts=full,ku_users AllowQos=ALL
   AllocNodes=ALL Default=NO QoS=N/A
   DefaultTime=UNLIMITED DisableRootJobs=NO ExclusiveUser=NO GraceTime=0 Hidden=NO
   MaxNodes=UNLIMITED MaxTime=UNLIMITED MinNodes=1 LLN=NO MaxCPUsPerNode=UNLIMITED MaxCPUsPerSocket=UNLIMITED
   Nodes=dgx-03
   PriorityJobFactor=1 PriorityTier=1 RootOnly=NO ReqResv=NO OverSubscribe=NO
   OverTimeLimit=NONE PreemptMode=OFF
   State=UP TotalCPUs=128 TotalNodes=1 SelectTypeParameters=NONE
   JobDefaults=(null)
   DefMemPerCPU=512 MaxMemPerNode=UNLIMITED
   TRES=cpu=128,mem=2064069M,node=1,billing=128,gres/gpu=8

PartitionName=gpuq
   AllowGroups=ALL AllowAccounts=root,mig,ku_users AllowQos=ALL
   AllocNodes=ALL Default=YES QoS=N/A
   DefaultTime=UNLIMITED DisableRootJobs=NO ExclusiveUser=NO GraceTime=0 Hidden=NO
   MaxNodes=UNLIMITED MaxTime=UNLIMITED MinNodes=1 LLN=NO MaxCPUsPerNode=UNLIMITED MaxCPUsPerSocket=UNLIMITED
   Nodes=dgx-[02,04]
   PriorityJobFactor=1 PriorityTier=1 RootOnly=NO ReqResv=NO OverSubscribe=NO
   OverTimeLimit=NONE PreemptMode=OFF
   State=UP TotalCPUs=384 TotalNodes=2 SelectTypeParameters=NONE
   JobDefaults=(null)
   DefMemPerCPU=512 MaxMemPerNode=UNLIMITED
   TRES=cpu=384,mem=4128104M,node=2,billing=384,gres/gpu=44,gres/gpu:1g.10gb=14,gres/gpu:2g.20gb=16,gres/gpu:3g.40gb=8,gres/gpu:7g.80gb=6
```

CPU: 18% MEM: 24GB NET: 2Gbps NONTRI-AI SERVER

NotebookLM

Jupyter Notebook





NotebookLM

ตัวอย่างงาน AI จริงบน HPC

Train Image Classification

Chat

User: “What is the status of the HPC job?”

LLM: “The job ‘ai-train’ is currently running on node g01 with 98.5% GPU utilization. Training epoch 5/100 is in progress.”

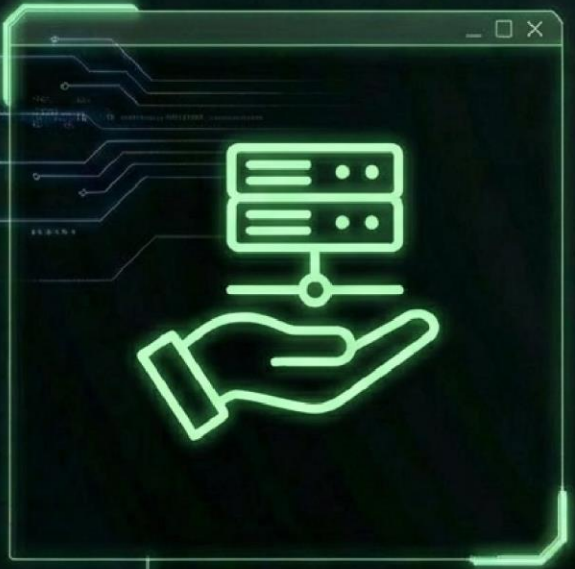
Fine-tune LLM

Data Processing ขนาดใหญ่

CPU: 18% MEM: 24GB NET: 2Gbps NONTRI-AI SERVER

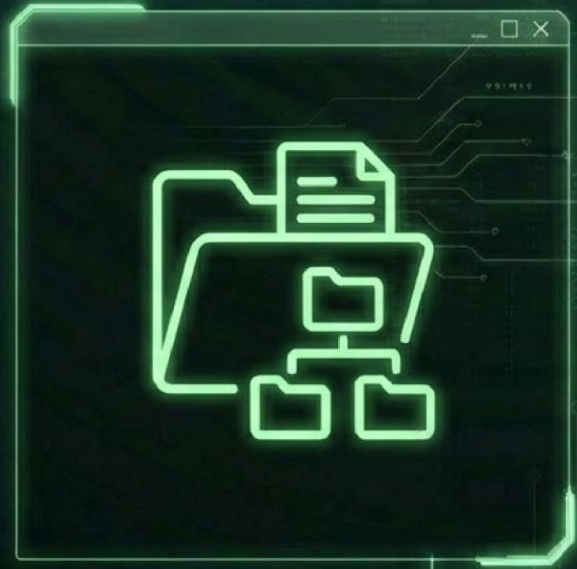
NotebookLM

สิ่งที่ต้องระวัง



- ใช้ทรัพยากรอย่างรับผิดชอบ

- ไม่รันงานทิ้งไว้ (Release resources)



- จัดการ Storage ให้เรียบร้อย

CPU: 18% MEM: 24GB NET: 2Gbps NONTRI-AI SERVER

NotebookLM


นี่คือสนามของคุณ
Nontri AI Gateway - โอกาสสู่ระดับที่เครื่องธรรมดาทำไม่ได้
CPU: 18% MEM: 24GB NET: 2Gbps NONTRI-AI SERVER

NotebookLM