eUne image contenant texte, capture d’écran, logiciel, Système d’exploitation

Description générée automatiquement

Sign up

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Email

Une image contenant texte, capture d’écran, conception

Description générée automatiquement

Une image contenant texte, logiciel, capture d’écran, Icône d’ordinateur

Description générée automatiquement

Menu > Billing

Choose amount

Button > Pay with card

Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Description générée automatiquement

Pay with your credit card (processed by Stripe)

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

… a few seconds later …

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

LINUX VM

ssh-keygen -t ed25519

cat ~/.ssh/id\_ed25519.pub

Menu > Settings

SSH public keys

Une image contenant texte, capture d’écran, logiciel, nombre

Description générée automatiquement

Button > Update public key

Une image contenant texte, logo, vert, Police

Description générée automatiquement

Menu > Pods

+ Deploy

Une image contenant texte, capture d’écran, logiciel, multimédia

Description générée automatiquement

Select a GPU – RTX4090

Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Description générée automatiquement

Scroll down

Buton > Change Template

Select “RunPod Pytorch 2.4.0 »

Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Description générée automatiquement

Check SSH terminal Access and Start Jupyter Notebook

Button > Edit Template

Select Disk size

Une image contenant texte, capture d’écran, logiciel, nombre

Description générée automatiquement

Button > Deploy On-Demand

… after 30 sec …

Une image contenant texte, capture d’écran, logiciel, Police

Description générée automatiquement

Click on the purple arrow to see pod details

Une image contenant texte, capture d’écran, logiciel, nombre

Description générée automatiquement

Button > Connect

Une image contenant texte, capture d’écran, logiciel, Page web

Description générée automatiquement

Button > Connect to Jupyter Lab

Une image contenant texte, logiciel, Icône d’ordinateur, Page web

Description générée automatiquement

Button > Start Web Terminal

Une image contenant texte, capture d’écran, Police, ligne

Description générée automatiquement

Button > Connect to Web Terminal

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Description générée automatiquement

Copy second SSH command to WSL command line

Une image contenant texte, capture d’écran, Police, diagramme

Description générée automatiquement

Button > Stop

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Pod status exited

Une image contenant texte, capture d’écran, logiciel, Police

Description générée automatiquement

Start or Terminate actions

Une image contenant capture d’écran, Police, Graphique, logo

Description générée automatiquement

Button > Edit Pod

Une image contenant texte, capture d’écran

Description générée automatiquement

You can change the disk sizes

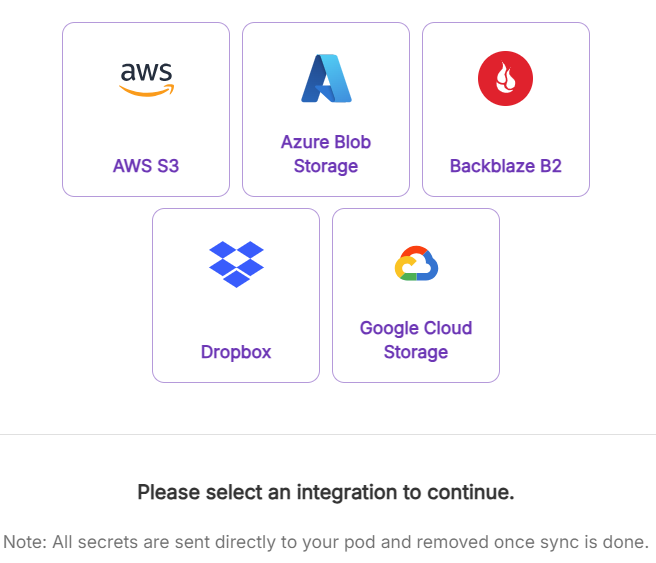
Button > Start Pod

Une image contenant texte, logiciel, Police, capture d’écran

Description générée automatiquement

You can choose 0 or 1 GPU, but YOU CAN’T CHANGE GPU

Button > Sync with cloud storage



INSIDE DOCKER CONTAINER

mfs#euro.runpod.net:9421 on /workspace type fuse (rw,nosuid,nodev,relatime,user\_id=0,group\_id=0,allow\_other)

/home and /workspace empty

/jupyter.log

Jupyterlab Version 4.2.5

[I 2024-10-21 21:13:18.843 LabApp] JupyterLab extension loaded from /usr/local/lib/python3.11/dist-packages/jupyterlab

[I 2024-10-21 21:13:18.843 LabApp] JupyterLab application directory is /usr/local/share/jupyter/lab

[I 2024-10-21 21:13:18.843 LabApp] Extension Manager is 'pypi'

Description: Ubuntu 22.04.5 LTS

/usr/bin/python

Python 3.11.10

/sbin/docker-init -- /opt/nvidia/nvidia\_entrypoint.sh /start.sh

/bin/bash /start.sh

/usr/bin/python /usr/local/bin/jupyter-lab --allow-root --no-browser --port=8888 --ip=\* --FileContentsManager.delete\_to\_trash=False --ServerApp.terminado\_settings

gotty -w --credential mr2ckkwvhh552kjdwcya:hyu4y1kqj6r9f1wzi0c8 --ws-origin .\* -p 19123 env TERM=xterm bash

/start.sh

# Start jupyter lab

start\_jupyter() {

if [[ $JUPYTER\_PASSWORD ]]; then

echo "Starting Jupyter Lab..."

mkdir -p /workspace && \

cd / && \

nohup jupyter lab --allow-root --no-browser --port=8888 --ip=\* --FileContentsManager.delete\_to\_trash=False --ServerApp.terminado\_settings='{"shell\_command":["/bin/bash"]}' --ServerApp.token=$JUPYTER\_PASSWORD --ServerApp.allow\_origin=\* --ServerApp.preferred\_dir=/workspace &> /jupyter.log &

echo "Jupyter Lab started"

fi

}

start\_nginx

execute\_script "/pre\_start.sh" "Running pre-start script..."

echo "Pod Started"

setup\_ssh

start\_jupyter

export\_env\_vars

execute\_script "/post\_start.sh" "Running post-start script..."

echo "Start script(s) finished, pod is ready to use."

sleep infinity

SET UP JupyterLab Git

pip install jupyterlab-git

ps aux

kill 79

nohup jupyter lab --allow-root --no-browser --port=8888 --ip=\* --FileContentsManager.delete\_to\_trash=False --ServerApp.terminado\_settings='{"shell\_command":["/bin/bash"]}' --ServerApp.token=$JUPYTER\_PASSWORD --ServerApp.allow\_origin=\* --ServerApp.preferred\_dir=/workspace &> /jupyter.log &

pip list

torch 2.4.1+cu124

(and nothing else)

apt list

build-essential / curl wget tar gzip

pip install torch==2.5 => 30 seconds

ENV VARIABLES

SHELL=/bin/bash

HOSTNAME=0da4ea99b944

PWD=/workspace

JUPYTER\_SERVER\_URL=http://localhost:8888/

RUNPOD\_CPU\_COUNT=8

HOME=/root

JUPYTER\_PASSWORD=ypwjv07zgi5uupb76w22

CUDA\_VERSION=12.4.1

RUNPOD\_POD\_ID=rx4dwyk56bhgq4

RUNPOD\_MEM\_GB=30

RUNPOD\_PUBLIC\_IP=213.173.98.80

RUNPOD\_VOLUME\_ID=h54ge2tjj4

RUNPOD\_GPU\_COUNT=1

JUPYTER\_SERVER\_ROOT=/

RUNPOD\_POD\_HOSTNAME=rx4dwyk56bhgq4-6441174d

RUNPOD\_DC\_ID=EU-RO-1

NVARCH=x86\_64

RUNPOD\_GPU\_NAME=NVIDIA+GeForce+RTX+4090

RUNPOD\_TCP\_PORT\_22=29463

RUNPOD\_API\_KEY=rpa\_JAD98YS2LWAMMW6FIWLVJABIH11X0F0OYHZMLL903l3gmb

PATH=/usr/local/nvidia/bin:/usr/local/cuda/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin

PUBLIC\_KEY=ssh-ed25519 AAAAC3NzaC1lZDI1NTE5AAAAIKtLXeyeiVlByoawRBWR3GUdU3OqpVl6WYmCZCwjtXqa root@Legion5Pro16IAH

OLDPWD=/

\_=/usr/bin/env

Network disk

mfs#euro.runpod.net:9421 on /workspace type fuse (rw,nosuid,nodev,relatime,user\_id=0,group\_id=0,allow\_other)

The volume disk is replaced by the network disk

* Only what is stored under /workspace is saved, everything else is lost

Select a network volume before starting a pod : it will limit the choice of GPUs to the same datacenter

Only option is to Terminate pod

When using a local volume : you first Stop pod, and then you can Start it again, but only with the same GPU, and finally if you Terminate the pod you loose all data

# New scripts



# Jarvislabs Analysis

Expose HTTP Ports : 6900,7860,8000,8080,8081,8888

ssh -o StrictHostKeyChecking=no -p 11114 [root@sshg.jarvislabs.ai](mailto:root@sshg.jarvislabs.ai)

git clone <https://github.com/wordslab-org/wordslab-notebooks.git>

cd wordslab-notebooks/server/

./0\_install\_on\_jarvislabs\_ai.sh

source ~./bashrc

PROBLEME DE VARIABLES ENV

git config --global user.email "laurent.prudhon@hotmail.fr"

git config --global user.name "Laurent Prud’hon"

git remote set-url origin https://ghp\_ABC123YOURTOKEN@github.com/wordslab-org/wordslab-notebooks.git

----

/home xfs

Ceph is a distributed storage platform that provides object storage, block storage, and file storage in a unified system. RBD (RADOS Block Devices) is the part of Ceph that provides block storage.

In cloud environments (such as OpenStack), RBDs are often used as the backend for virtual machine disk storage.

XFS doesn’t support shrinking a file system (you can extend the file system, but not reduce its size once it's allocated).

everything else Docker overlay

File storage

/dev/rbd2 on /home type xfs (rw,relatime,attr2,inode64,logbufs=8,logbsize=64k,sunit=128,swidth=128,noquota)

admin@e146475a-ca58-11ee-991b-456cac8ab788.jl\_usr\_fs=/ on /home/jl\_fs type ceph (rw,relatime,name=admin,secret=<hidden>,ms\_mode=prefer-crc,acl,mon\_addr=10.232.204.13:3300/10.232.204.18:3300/10.232.204.19:3300/10.232.204.21:3300/10.232.204.22:3300)

Ssh connexion

ssh -o StrictHostKeyChecking=no -p 11214 root@sshg.jarvislabs.ai

which python

/root/miniconda3/envs/py3.10/bin/python

Ubuntu 22.04.4 LTS

CUDA Version: 12.4

/root/miniconda3/envs/py3.10/bin/python

/root/miniconda3/envs/py3.10/lib/python3.10/site-packages

torch 2.4.0

jupyterlab 4.2.4

diffusers 0.30.0

huggingface\_hub 0.24.5

transformers-4.44.0

pandas 2.2.2

spacy 3.7.5

env

MACHINE\_ID=214002

PATH=/root/miniconda3/envs/py3.10/bin:/root/miniconda3/bin:/usr/local/nvidia/bin:/usr/local/cuda/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin

GRADIO\_SERVER\_PORT=6006

GRADIO\_SERVER\_NAME=0.0.0.0

MACHINE\_NAME=distill-whisper-test

STREAMLIT\_SERVER\_PORT=6006

JUPYTER\_SERVER\_ROOT=/home

TOKEN=537\_-uA2Ps751OGNxYbunQ09ewihBG4t\_tYH46XpZd\_3k-lFqCfZZES3NyTqhWuL

JUPYTER\_SERVER\_URL=http://fd17b30a3057:8889/

HOSTNAME=fd17b30a3057

PYTHON\_VERSION=3.10

USER\_ID=laurent.prudhon@hotmail.fr

if [ -f /.dockerenv ]; then

echo "Running inside a Docker container"

else

echo "Not running inside a Docker container"

fi

root@fd17b30a3057:~# conda env list

# conda environments:

#

base /root/miniconda3

py3.10 /root/miniconda3/envs/py3.10

root@fd17b30a3057:~# cat /docker-entrypoint.sh

#!/bin/bash

source /opt/conda/etc/profile.d/conda.sh

conda activate py3.10

echo "PasswordAuthentication no" >> /etc/ssh/sshd\_config

service ssh start

export SHELL="/bin/bash"

env HOME=/home code-server --host 0.0.0.0 --port 7007 --auth none&

env HOME=/home jupyter lab --ip=0.0.0.0 --NotebookApp.token=$TOKEN --allow-root --port 8889

root 1 0.0 0.0 1136 0 ? Ss 15:59 0:00 /sbin/docker-init -- /docker-entrypoint.sh

root 7 0.0 0.0 4364 2048 ? S 15:59 0:00 /bin/bash /docker-entrypoint.sh

root 33 0.0 0.0 15432 3108 ? Ss 15:59 0:00 sshd: /usr/sbin/sshd [listener] 0 of 10-100 startups

root 34 0.0 0.0 1386032 63232 ? Sl 15:59 0:00 /usr/lib/code-server/lib/node /usr/lib/code-server --host 0.0.0.0 --port 7007 --auth none

root 35 0.2 0.0 892632 101924 ? Rl 15:59 0:13 /root/miniconda3/envs/py3.10/bin/python /root/miniconda3/envs/py3.10/bin/jupyter-lab --ip=0.0.0.0 --NotebookApp.token=537\_-uA2Ps751OGNxYbunQ09ewihBG4t\_tYH46XpZd\_3k-lF

root 53 0.0 0.0 1321328 64276 ? Sl 15:59 0:00 /usr/lib/code-server/lib/node /usr/lib/code-server/out/node/entry

root 70 0.0 0.0 5004 3072 pts/0 Ss 16:00 0:00 /bin/bash -l

root 82 0.0 0.0 1377688 57344 ? Ssl 16:00 0:00 /root/miniconda3/envs/py3.10/bin/python -m ipykernel\_launcher -f /home/.local/share/jupyter/runtime/kernel-c8239b0f-5830-43f6-8b48-20e93c7a63cc.json

/home/start\_up.sh

=> executed after all is launched

20 GB | $0.0028/hr => "Jarvislabs prices.xlsx" sur le bureau

1 CPU 2GB RAM => 0,02 Euros

A5000 24GB VRAM, 32 CPUs 64GB RAM = RTX 3080 with VRAM x2 => 0,49 Euros

RTX6000Ada 48GB VRAM, 32 CPUs 128GB RAM = RTX 4090 with VRAM x2 => 0,99 Euros

You should have only one instance

100 GB for 10$ per month

Manage the space on your instance

Launch with CPU when you want to download/install or develop => virtually free

Launch with GPU when you want to test/generate/train

Local instance mandatory

But CPU version available

Remote instance is set up from a local notebook

Include ssh key generation in local install

ssh-keygen -t ed25519

cat ~/.ssh/id\_ed25519.pub

https://jarvislabs.ai/docs/settings/setup-ssh

When clicking on the SSH logo

ssh -o StrictHostKeyChecking=no -p 21414 root@sshb.jarvislabs.ai

conda init

source ~/.bashrc

conda activate py3.10

conda config --add channels conda-forge

conda install -y jupyterlab\_execute\_time=3.2.0

conda install -y jupyterlab-nvdashboard=0.11.00

conda install -y jupyterlab-git=0.50.1

conda install -y ipympl=0.9.4

Configure Auto Pause as a security

https://jarvislabs.ai/docs/api

https://jarvislabs.ai/docs/api#pausing-inside-a-jarvislabs-instance

# Scripts design

Current focus = NVIDIA GPUs / build AI systems, not inference only

* 1. Ubuntu 22.04+ Linux machine (VM or container) with remote access (wsl or ssh)
  2. Install and configure Ubuntu packages to ensure a common ground  
     => docker support only if not inside container
  3. Configure persistent root storage paths : primary disk, shared disk (can point to the same)  
     => WORKSPACE and MODELS
  4. Configure open ports for several APÏs & UIs (LAN or internet access)  
     Jupyterlab, Gradio, fastapi & fasthtml & VLLM, argilla.io, Open WebUI, VS Code server  
     $defaultports = 8888, 7860, 8000, 6900, 8080, 8081

2.1 Python environnement for deep learning with Pytorch 2.4  
=> raw install or conda environnement  
=> GPU version with CUDA 12.4 or CPU version with Intel MKL  
2.2 Two layers for Python librairies : scripts to create virtual environments to install the latest frameworks (in persistent location)   
2.3 Configuration for models download in a persistent location

3.1 Jupyterlab 4.2+ and VsCode  
3.2 Git Support  
3.3 Resources monitoring support