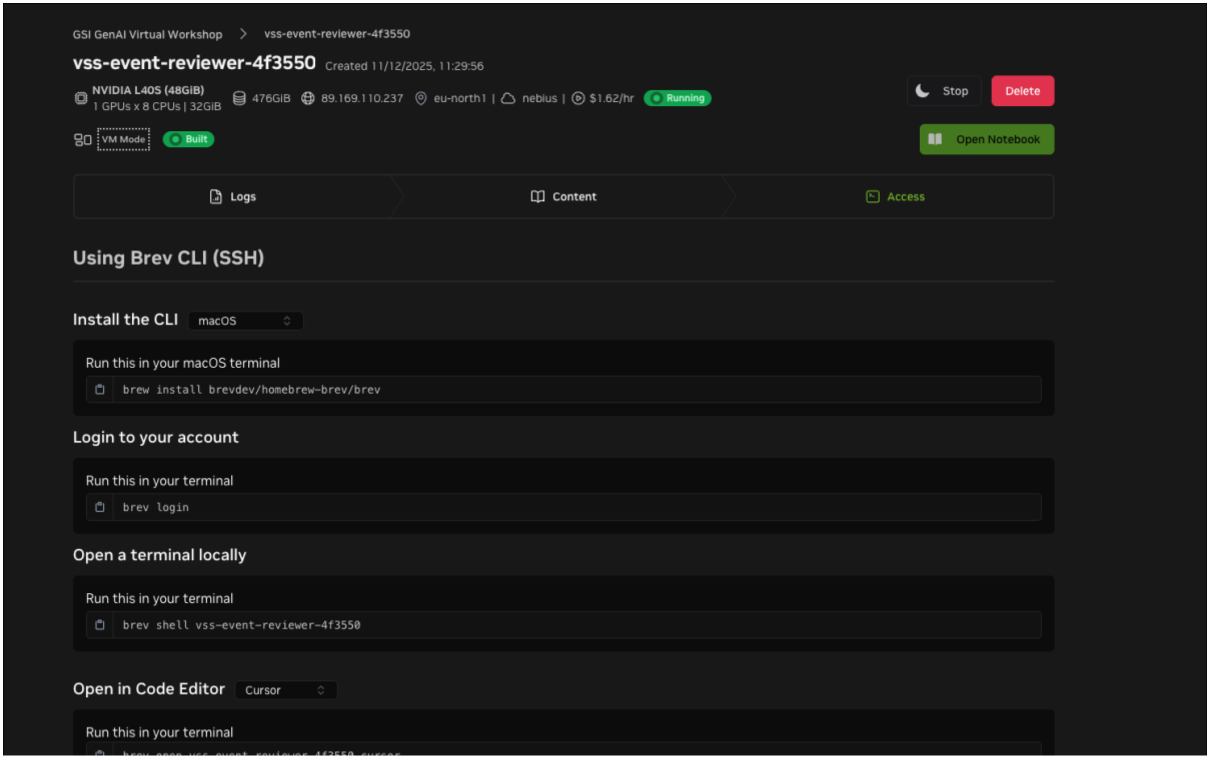
Steps to setup the cluster using docker compose

1. Launch the instance through brev UI using the launchable for vss.



GSI GenAI Virtual Workshop > vss-event-reviewer-4f3550

vss-event-reviewer-4f3550 Created 11/12/2025, 11:29:56

NVIDIA L40S (48GiB)
1 GPUs x 8 CPUs | 32GiB 476GiB 89.169.110.237 eu-north1 | nebius | $1.62/hr Running

VM Mode Built

Using Brev CLI (SSH)

Install the CLI macOS

Run this in your macOS terminal

```
brew install brevdev/homebrew-brev/brev
```

Login to your account

Run this in your terminal

```
brev login
```

Open a terminal locally

Run this in your terminal

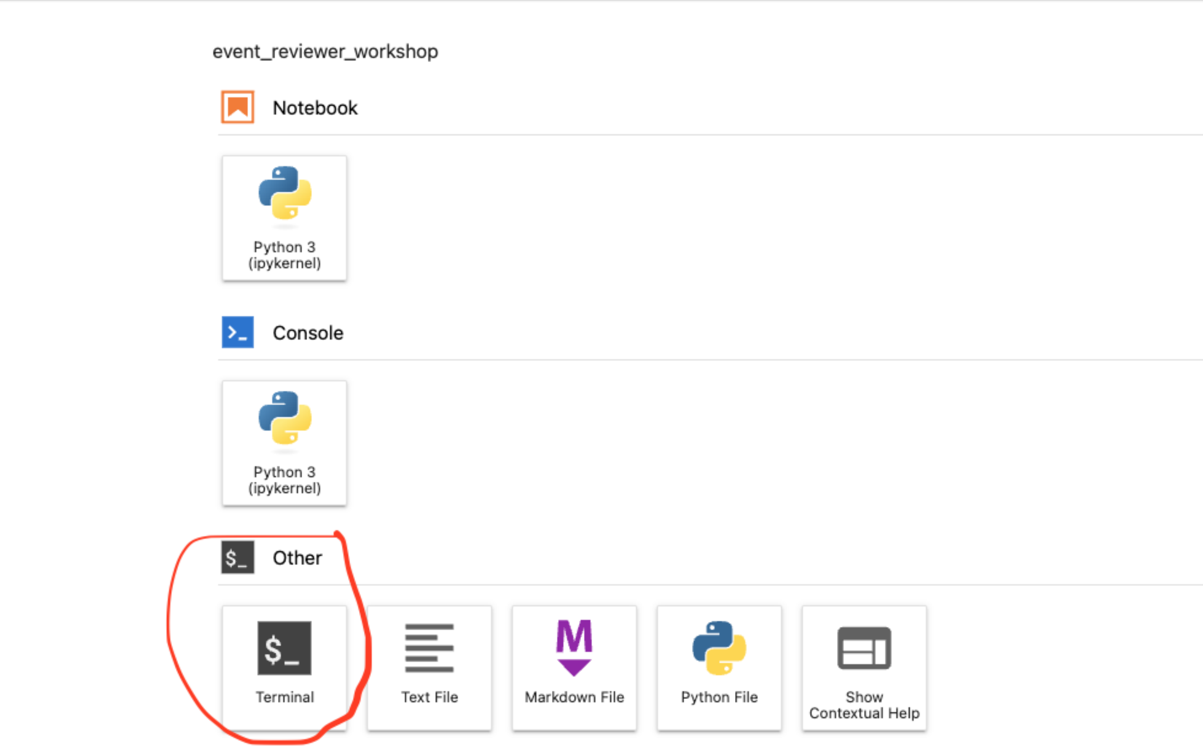
```
brev shell vss-event-reviewer-4f3550
```

Open in Code Editor Cursor

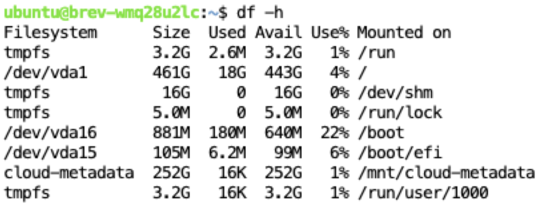
Run this in your terminal

```
brev open vss-event-reviewer-4f3550 cursor
```

2. Open the terminal through the jupyter launcher and proceed through the following steps(highlighted in red):

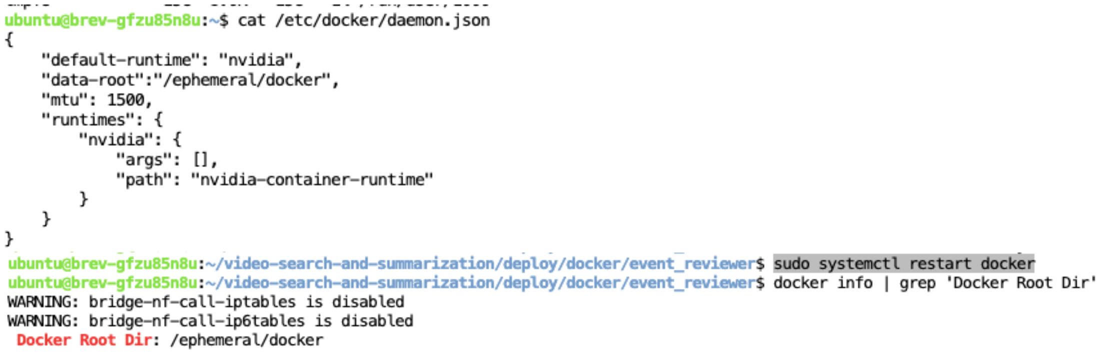


3. Check the file storage space using **‘df -h’** in a terminal window on the jupyter notebook - to verify if sufficient storage is available in the root directory “/”



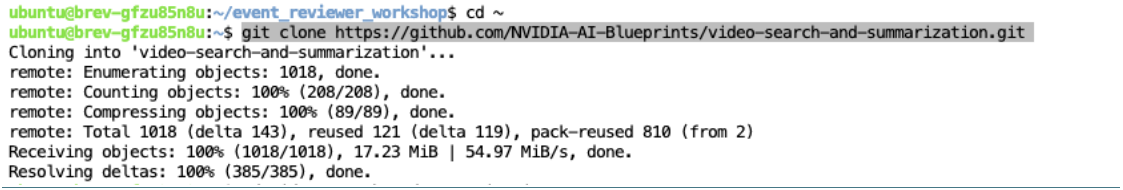
```
ubuntu@brev-wmq28u2lc:~$ df -h
Filesystem      Size  Used Avail Use% Mounted on
tmpfs           3.2G  2.6M  3.2G   1% /run
/dev/vda1       461G   18G  443G   4% /
tmpfs            16G     0   16G   0% /dev/shm
tmpfs           5.0M     0  5.0M   0% /run/lock
/dev/vda16      881M  180M  640M  22% /boot
/dev/vda15      105M  6.2M   99M   6% /boot/efi
cloud-metadata  252G   16K  252G   1% /mnt/cloud-metadata
tmpfs           3.2G   16K  3.2G   1% /run/user/1000
```

4. [Optional] For cloud providers like CRUSOE, the main data mount is stored in a path like “/ephermal”. Add this path to the “/etc/docker/daemon.json” using ‘*sudo vi /etc/docker/daemon.json’* and restart docker *‘sudo systemctl restart docker’*



```
ubuntu@brev-gfzu85n8u:~$ cat /etc/docker/daemon.json
{
    "default-runtime": "nvidia",
    "data-root":"/ephemeral/docker",
    "mtu": 1500,
    "runtimes": {
        "nvidia": {
            "args": [],
            "path": "nvidia-container-runtime"
        }
    }
}
ubuntu@brev-gfzu85n8u:~/video-search-and-summarization/deploy/docker/event_reviewer$ sudo systemctl restart docker
ubuntu@brev-gfzu85n8u:~/video-search-and-summarization/deploy/docker/event_reviewer$ docker info | grep 'Docker Root Dir'
WARNING: bridge-nf-call-iptables is disabled
WARNING: bridge-nf-call-ip6tables is disabled
 Docker Root Dir: /ephemeral/docker
```

5. Go to the parent directory ‘**cd ~**’ and git clone the VSS respository ‘**git clone [https://github.com/NVIDIA-AI-Blueprints/video-search-and-summarization.git’](https://github.com/NVIDIA-AI-Blueprints/video-search-and-summarization.git)**

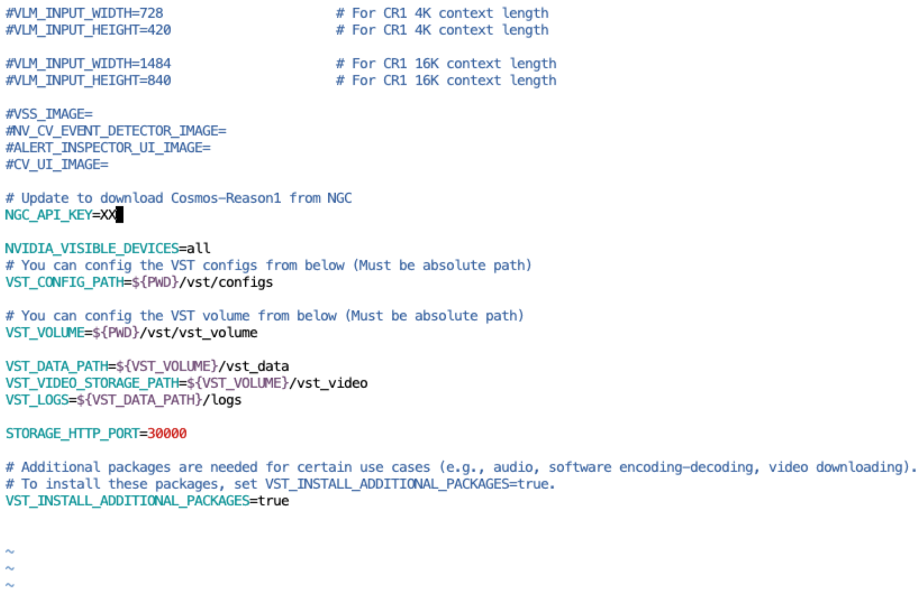


```
ubuntu@brev-gfzu85n8u:~/event_reviewer_workshop$ cd ~
ubuntu@brev-gfzu85n8u:~$ git clone https://github.com/NVIDIA-AI-Blueprints/video-search-and-summarization.git
Cloning into 'video-search-and-summarization'...
remote: Enumerating objects: 1018, done.
remote: Counting objects: 100% (208/208), done.
remote: Compressing objects: 100% (89/89), done.
remote: Total 1018 (delta 143), reused 121 (delta 119), pack-reused 810 (from 2)
Receiving objects: 100% (1018/1018), 17.23 MiB | 54.97 MiB/s, done.
Resolving deltas: 100% (385/385), done.
```

6. Move into the folder ‘**cd ~/video-search-and-summarization/deploy/docker/event_reviewer**’

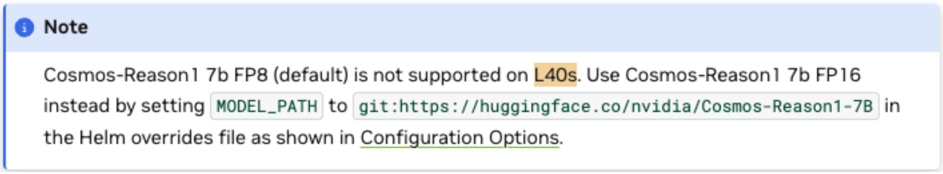
7. Use NGC API key from section Obtain NGC API Key.
   Update NGC_API_KEY environment variable in .env file in the ‘./**video-search-and-**

summarization/deploy/docker/event_reviewer' folder to a valid key.



```
#VLM_INPUT_WIDTH=728                      # For CR1 4K context length
#VLM_INPUT_HEIGHT=420                     # For CR1 4K context length

#VLM_INPUT_WIDTH=1484                     # For CR1 16K context length
#VLM_INPUT_HEIGHT=840                     # For CR1 16K context length

#VSS_IMAGE=
#NV_CV_EVENT_DETECTOR_IMAGE=
#ALERT_INSPECTOR_UI_IMAGE=
#CV_UI_IMAGE=

# Update to download Cosmos-Reason1 from NGC
NGC_API_KEY=XX

NVIDIA_VISIBLE_DEVICES=all
# You can config the VST configs from below (Must be absolute path)
VST_CONFIG_PATH=${PWD}/vst/configs

# You can config the VST volume from below (Must be absolute path)
VST_VOLUME=${PWD}/vst/vst_volume

VST_DATA_PATH=${VST_VOLUME}/vst_data
VST_VIDEO_STORAGE_PATH=${VST_VOLUME}/vst_video
VST_LOGS=${VST_DATA_PATH}/logs

STORAGE_HTTP_PORT=30000

# Additional packages are needed for certain use cases (e.g., audio, software encoding-decoding, video downloading).
# To install these packages, set VST_INSTALL_ADDITIONAL_PACKAGES=true.
VST_INSTALL_ADDITIONAL_PACKAGES=true


~
~
~
```

8. For running on L40S update the model path to git:https://huggingface.co/nvidia/Cosmos-Reason1-7B as default FP8 is not supported on L40S.

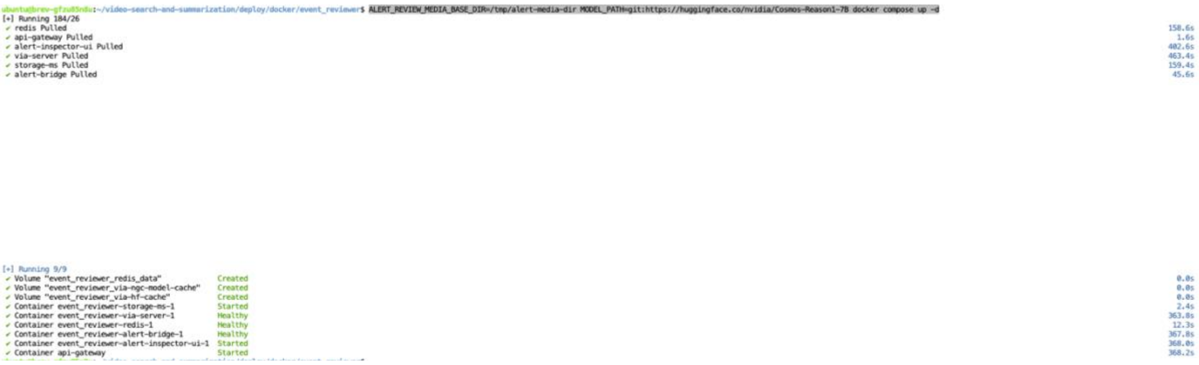


Note

Cosmos-Reason1 7b FP8 (default) is not supported on L40s. Use Cosmos-Reason1 7b FP16 instead by setting `MODEL_PATH` to `git:https://huggingface.co/nvidia/Cosmos-Reason1-7B` in the Helm overrides file as shown in Configuration Options.

9. Run the command as '**ALERT_REVIEW_MEDIA_BASE_DIR=/tmp/alert-media-dir MODEL_PATH=git:https://huggingface.co/nvidia/Cosmos-Reason1-7B docker compose up -d**'

10. Change permissions of the '/tmp/alert-media-dir' to 777 using '**sudo chmod 777 /tmp/alert-media-dir/**'



```
ubuntu@brev-gfzu85n8u:~/video-search-and-summarization/deploy/docker/event_reviewer$ ALERT_REVIEW_MEDIA_BASE_DIR=/tmp/alert-media-dir MODEL_PATH=git:https://huggingface.co/nvidia/Cosmos-Reason1-7B docker compose up -d
[+] Running 184/26
 ✔ redis Pulled                                                                   158.6s
 ✔ api-gateway Pulled                                                               1.6s
 ✔ alert-inspector-ui Pulled                                                      402.6s
 ✔ via-server Pulled                                                              463.4s
 ✔ storage-ms Pulled                                                              159.4s
 ✔ alert-bridge Pulled                                                             45.6s

[+] Running 9/9
 ✔ Volume "event_reviewer_redis_data"              Created                          0.0s
 ✔ Volume "event_reviewer_via-ngc-model-cache"     Created                          0.0s
 ✔ Volume "event_reviewer_via-hf-cache"            Created                          0.0s
 ✔ Container event_reviewer-storage-ms-1           Started                          2.4s
 ✔ Container event_reviewer-via-server-1           Healthy                        363.8s
 ✔ Container event_reviewer-redis-1                Healthy                         12.3s
 ✔ Container event_reviewer-alert-bridge-1         Healthy                        367.8s
 ✔ Container event_reviewer-alert-inspector-ui-1   Started                        368.0s
 ✔ Container api-gateway                           Started                        368.2s
```

11. Follow the documentation here https://docs.nvidia.com/vss/latest/content/vss_event_reviewer.html#starting-the-deployment for more details

12.