Google Developer Groups · Cloud Chennai

cloud community days 2025

12 July 2025

Treat Your AI Models Like You Treat Containers

Ram Iyengar

Models

Main Tasks Libraries Languages Licenses Other

Tasks

Text Generation Any-to-Any Image-Text-to-Text
Image-to-Text Image-to-Image Text-to-Image
Text-to-Video Text-to-Speech + 43

Parameters

< 1B 6B 12B 32B 128B > 500B

Libraries

PyTorch TensorFlow JAX Transformers
Diffusers Safetensors ONNX GGUF
Transformers.js MLX MLX Keras + 40

Apps

vLLM TGI llama.cpp MLX LM
LM Studio Ollama Jan + 12

Inference Providers

Fireworks Cerebras Novita Featherless AI
Nebius AI Together AI SambaNova
Hyperbolic + 6

Models 1,858,939 Filter by name Full-text search ↑↓ Sort: Trending

THUDM/GLM-4.1V-9B-Thinking
Image-Text-to-Text • 10B • Updated 3 days ago • 24.3k • 557

HuggingFaceTB/SmolLM3-3B
Text Generation • 3B • Updated 1 day ago • 17k • 337

moonshotai/Kimi-K2-Instruct
Text Generation • Updated about 10 hours ago • 12.2k • 275

black-forest-labs/FLUX.1-Kontext-dev
Image-to-Image • Updated 14 days ago • 217k • 1.55k

apple/DiffuCoder-7B-cpGRPO
8B • Updated 7 days ago • 3.06k • 275

mistralai/Devstral-Small-2507
Text Generation • 24B • Updated 1 day ago • 1.48k • 142

ChatDOC/OCRFlux-3B
Image-to-Text • 4B • Updated 3 days ago • 8.95k • 269

kyutai/tts-1.6b-en_fr
Text-to-Speech • Updated 3 days ago • 23.8k • 275

google/gemma-3n-E4B-it
Image-Text-to-Text • 8B • Updated 1 day ago • 256k • 541

microsoft/Phi-4-mini-flash-reasoning
Text Generation • 4B • Updated about 20 hours ago • 587 • 103

LiquidAI/LFM2-1.2B
Text Generation • 1B • Updated about 9 hours ago • 842 • 102

tngtech/DeepSeek-TNG-R1T2-Chimera
Text Generation • 685B • Updated 1 day ago • 3.86k • 197

HuggingFaceTB/SmolLM3-3B-Base
Text Generation • 3B • Updated about 14 hours ago • 2.52k • 92

moonshotai/Kimi-K2-Base
Text Generation • Updated about 10 hours ago • 3 • 90

nanonets/Nanonets-OCR-s
Image-Text-to-Text • 4B • Updated 21 days ago • 293k • 1.38k

black-forest-labs/FLUX.1-dev
Text-to-Image • Updated 14 days ago • 1.52M • 10.8k

KurmaAI/AQUA-7B
Text Generation • 7B • Updated 6 days ago • 686 • 67

deepseek-ai/DeepSeek-R1-0528
Text Generation • 685B • Updated May 29 • 260k • 2.23k

tencent/Hunyuan-A13B-Instruct
Text Generation • 80B • Updated 4 days ago • 31.6k • 759

NovelAI/nai-anime-v2
Text-to-Image • Updated 6 days ago • 733 • 57

meta-llama/Llama-3.1-8B-Instruct
Text Generation • 8B • Updated Sep 25, 2024 • 5.5M • 4.28k

google/medgemma-27b-it
Image-Text-to-Text • 29B • Updated 2 days ago • 438 • 55

jinaai/jina-embeddings-v4
Visual Document Retrieval • 4B • Updated 2 days ago • 26.2k • 230

RekaAI/reka-flash-3.1
21B • Updated 2 days ago • 141 • 53

agentica-org/DeepSWE-Preview
Text Generation • 33B • Updated 9 days ago • 3.37k • 143

google/gemma-3n-E4B-it-litert-preview
Image-Text-to-Text • Updated May 26 • 1.36k

vrgamedevgirl84/Wan14BT2VFusioniX
Text-to-Video • Updated 20 days ago • 412

LiquidAI/LFM2-350M
Text Generation • 0.4B • Updated about 9 hours ago • 156 • 51

Skywork/Skywork-R1V3-38B

K-intelligence/Midm-2.0-Base-Instruct

Containers

Unified packaging

Portable

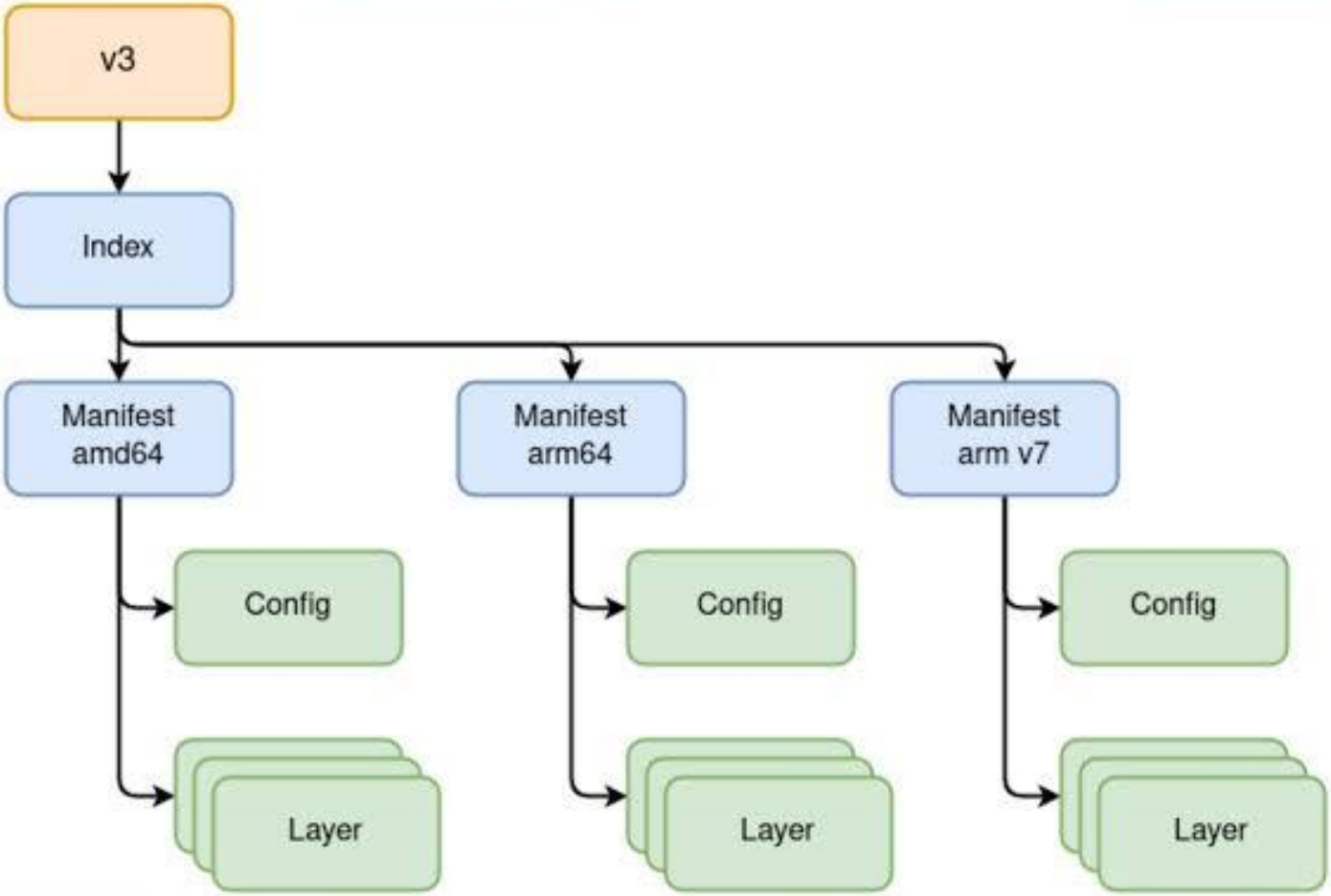
Worked for all languages

Fragmented

Opinionated


Let’s create
~~a scandal~~
a standard!

OCI Artifacts


v3
Index
Manifest
amd64
Manifest
arm64
Manifest
arm v7
Config
Config
Config
Layer
Layer
Layer

Content addressable

Group related items

AI/ML

Data in DvC

Technical debt, MLOps tools

Code in git repos

Config in feature stores

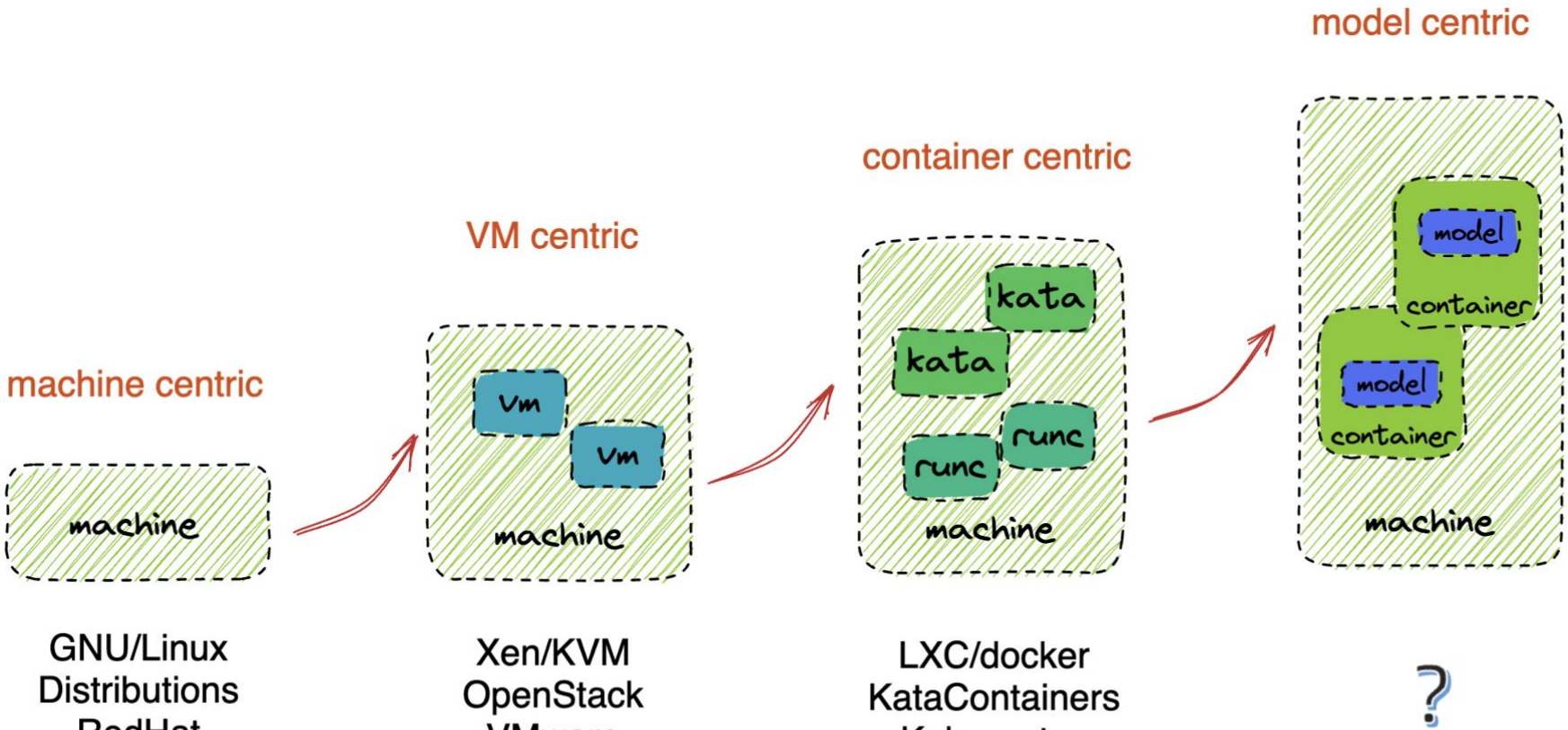
Data in S3

Pipeline definitions ...

Code in Jupyter notebooks

Fragmented

Opinionated


machine centric
machine
GNU/Linux
Distributions
RedHat
SuSE
VM centric
vm
vm
machine
Xen/KVM
OpenStack
VMware
AWS
container centric
kata
kata
runc
runc
machine
LXC/docker
KataContainers
Kubernetes
CNCF
model centric
model
container
model
container
machine
?

Docker

Reproducible

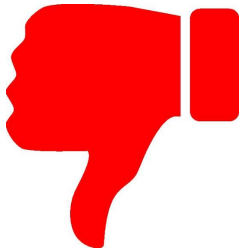
Reuse existing pipelines

Filesystems are layered

Not AI-native

Not content addressable

Perf, security, and efficiency



MLFlow

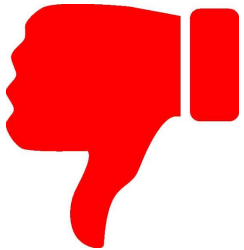
Tracking

Projects

Models, Model Registry

Lacked basic scaling

Ops complexity



Kserve

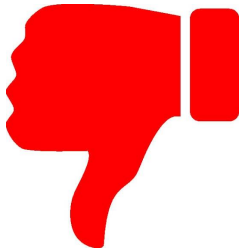
CRD based automation

Multiple-step inference

Serverless AI/MLOps

No advance customization

Batch inference



KubeRay

Kubernetes Operators/CRD

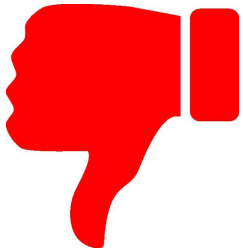
RayTrain, RayServe, RayTune

Serverless AI/MLOps

Complexity

Re-architect everything

Debugging 🤯



ONNX

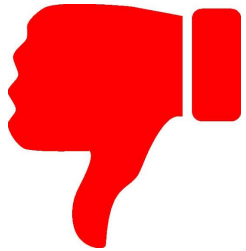
First known open standard

Converted models to .onnx

Boosts speed, compresses

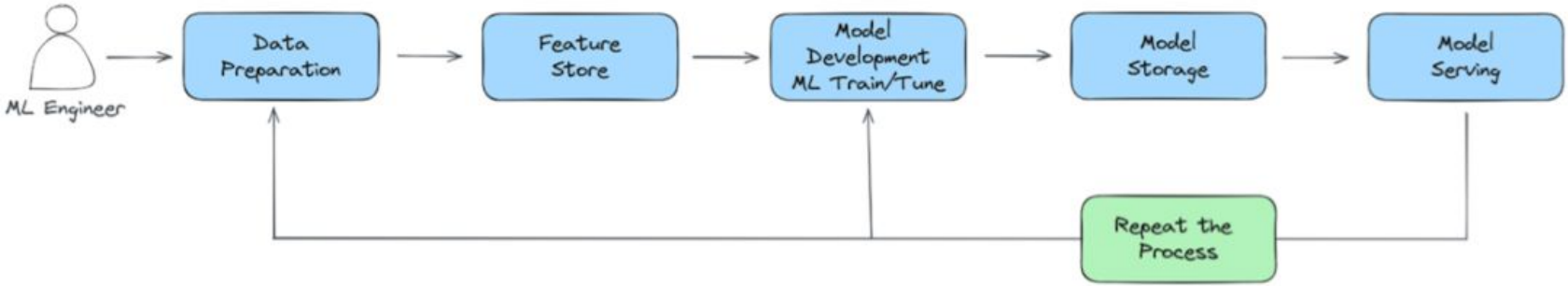
Tuning for ONNX support

New architecture support

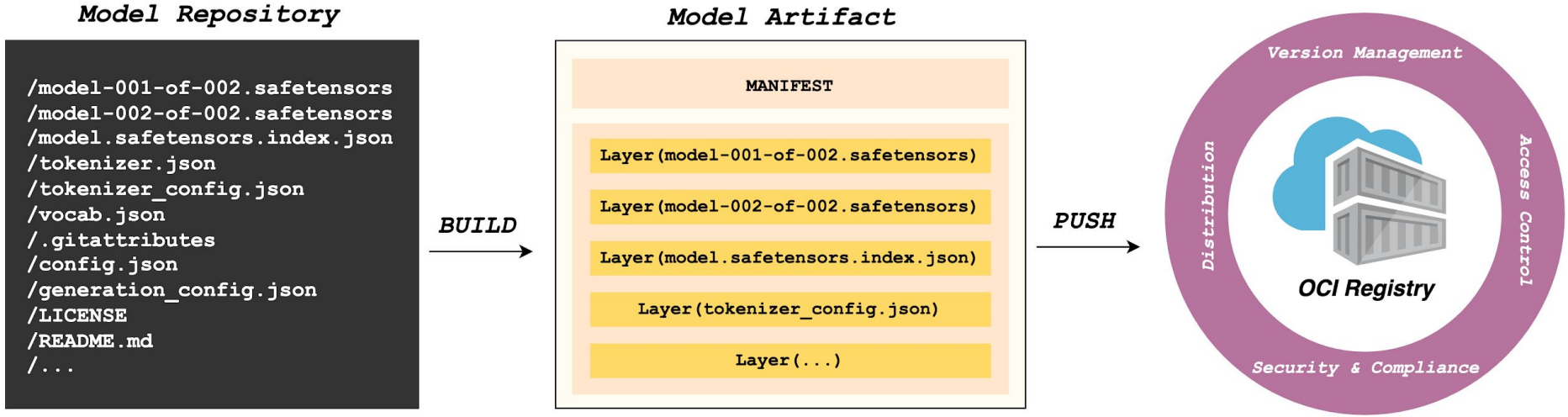


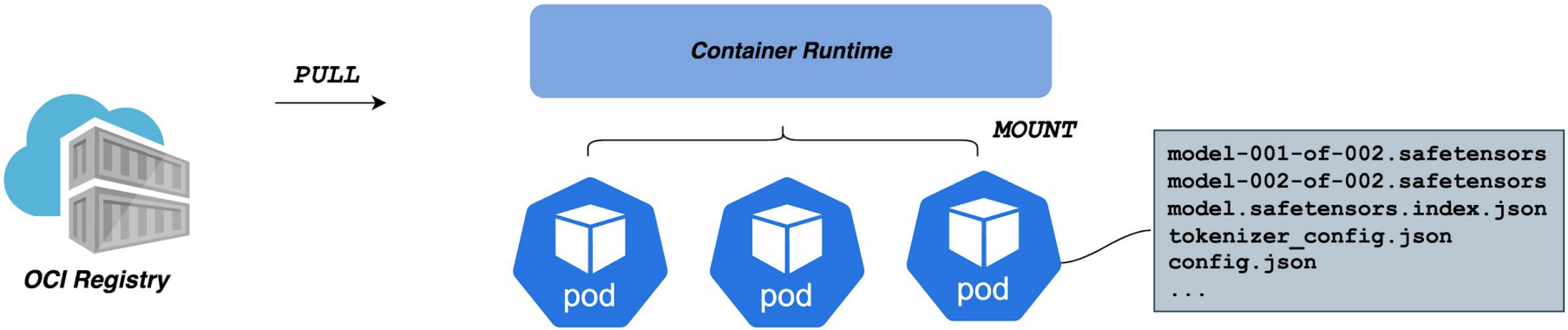
KitOps

Open ModelPack Spec



ML Lifecycle


Model Repository
/model-001-of-002.safetensors
/model-002-of-002.safetensors
/model.safetensors.index.json
/tokenizer.json
/tokenizer_config.json
/vocab.json
/.gitattributes
/config.json
/generation_config.json
/LICENSE
/README.md
/...
BUILD
Model Artifact
MANIFEST
Layer(model-001-of-002.safetensors)
Layer(model-002-of-002.safetensors)
Layer(model.safetensors.index.json)
Layer(tokenizer_config.json)
Layer(...)
PUSH
Version Management
Access Control
Security & Compliance
Distribution
OCI Registry


PULL
Container Runtime
MOUNT
OCI Registry
pod
pod
pod
model-001-of-002.safetensors
model-002-of-002.safetensors
model.safetensors.index.json
tokenizer_config.json
config.json
...

ModelKit

@ramiyengar