



**KubeCon**



**CloudNativeCon**

**North America 2024**





KubeCon



CloudNativeCon

North America 2024

# Managing and Distributing AI Models Using OCI Standards and Harbor

Steven Zou / Software Engineer @VMW by Broadcom | Harbor Maintainer  
Steven Ren/ Engineering Director @VMW by Broadcom | Harbor Maintainer

# Agenda

- Motivations
- Background
- Model Registry Idea
- Implementation atop Harbor
- Demos
- What's Next?

# Motivations

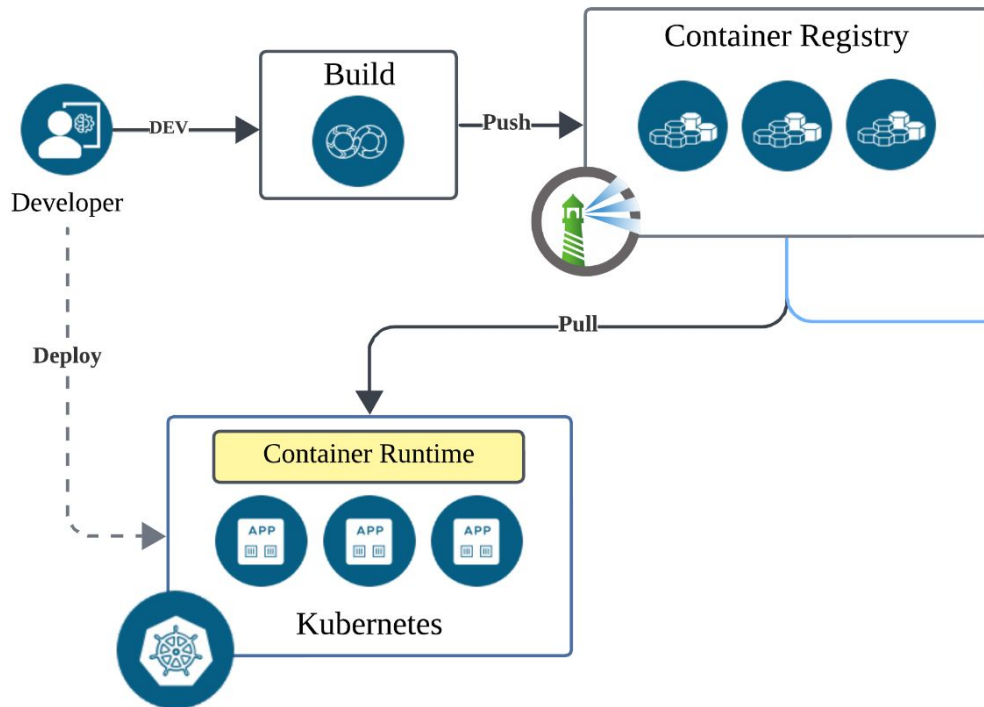


KubeCon



CloudNativeCon

North America 2024



**Kubernetes:** Platform for container orchestration and workload management

**Container Image Registry:** container distribution and deployment; key to the success.

**AI Workload:** code + model, dataset

**Private AI model registry service is a hard requirement.**

---



Intelligent  
Property  
Protection



Security &  
Compliance



Performance

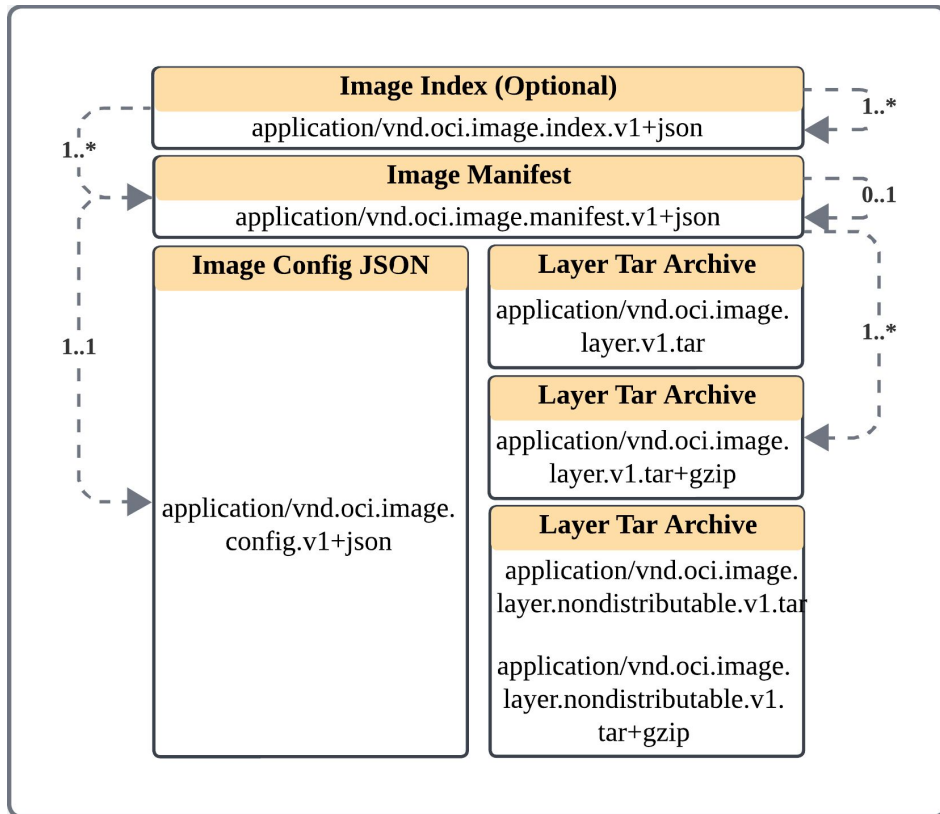


Cost  
Efficiency

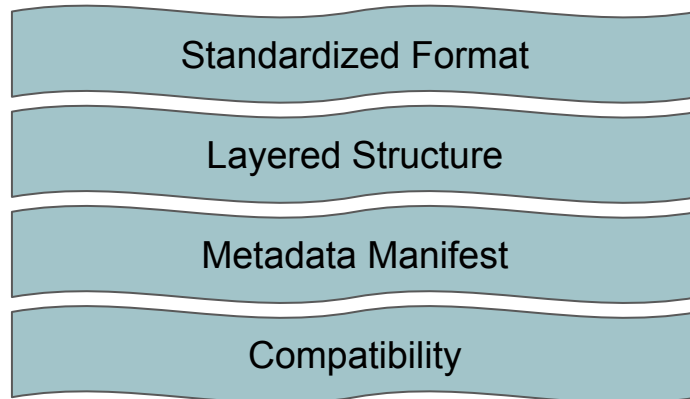


Smooth  
Integration

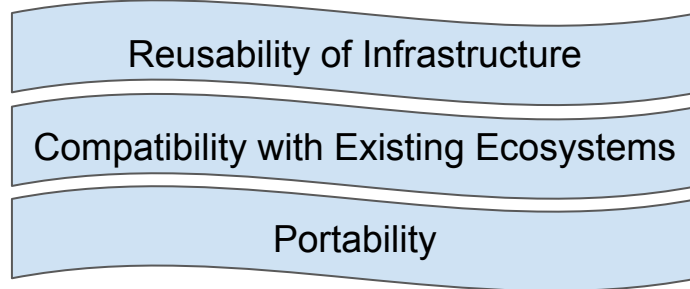
## OCI Image Spec



## Characteristics :

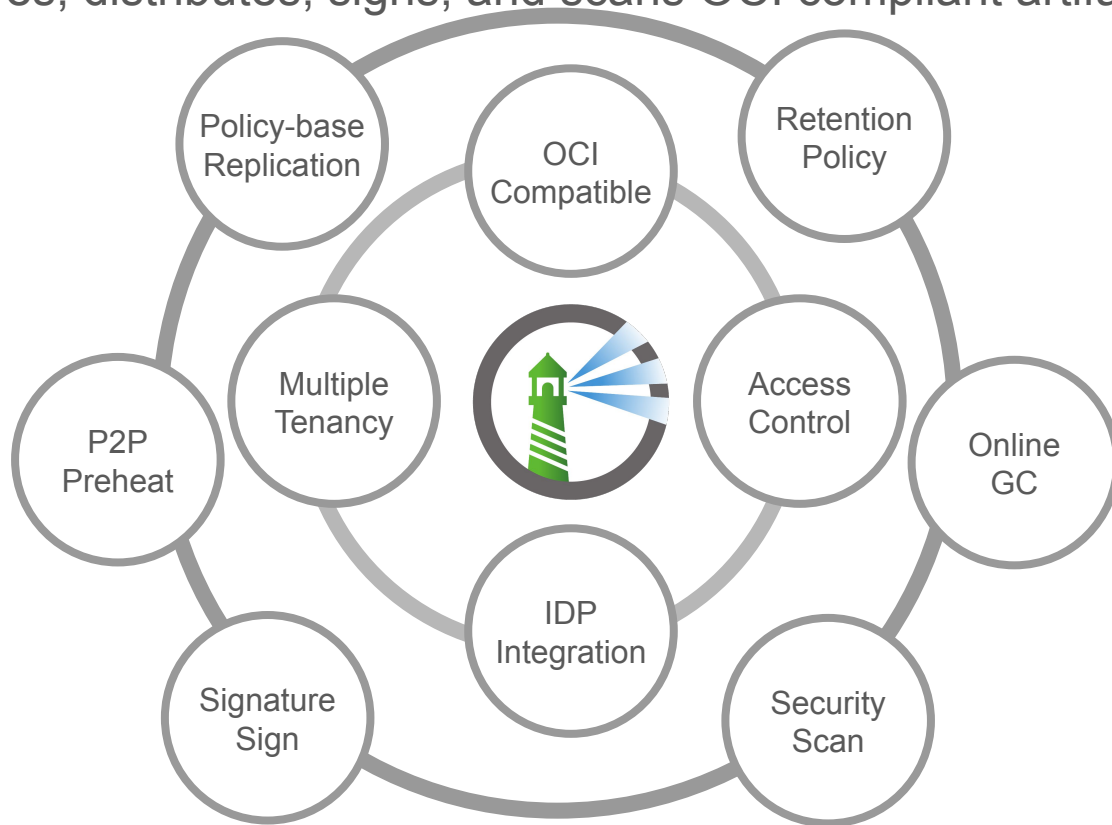


## Benifits :



# Background - Harbor

Harbor, a CNCF-graduated, open-source trusted cloud native registry project; stores, distributes, signs, and scans OCI compliant artifacts.



# History Related to AI/ML Model

Harbor has a long history to start the AI model Journey.

## Proposal for enhanced default processor to handle user-defined artifact like Machine Learning artifacts #143

Merged steven-zou merged 16 commits into goharbor:master from hyy0322:artifact-processor-extender on Aug 27, 2020

Conversation 104 Commits 16 Checks 0 Files changed 8 +316 -0



hyy0322 commented on May 31, 2020 · edited

Contributor

Signed-off-by: Yiyang Huang [huangyiyang.huangyy@bytedance.com](mailto:huangyiyang.huangyy@bytedance.com)

4 5

hyy0322 force-pushed the artifact-processor-extender branch 3 times, most recently from 33e62b1 to 258bda8 4 years ago

Compare

gaocegege mentioned this pull request on May 31, 2020

[feature request] Extend OCI Artifact Types in Runtime goharbor/harbor#12013

Closed

Reviewers

reasonerjt

gaocegege

ywk253100

zhujian7

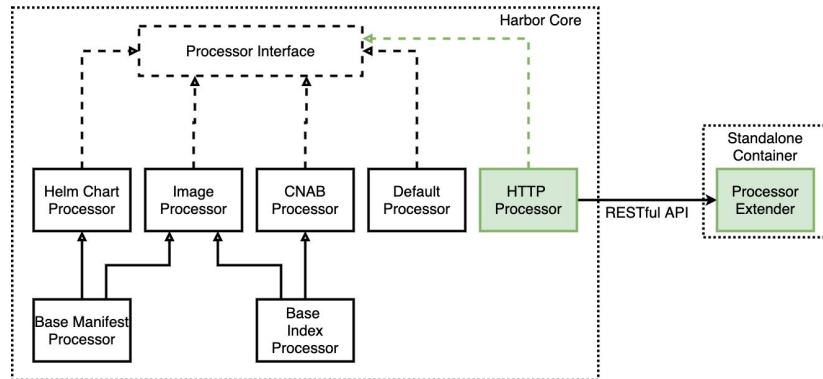
xaleeks

steven-zou

Assignees

No one—assign yourself

Collaborate with Harbor community partner CaiCloud (aka. ByteDance Volcano Engine)



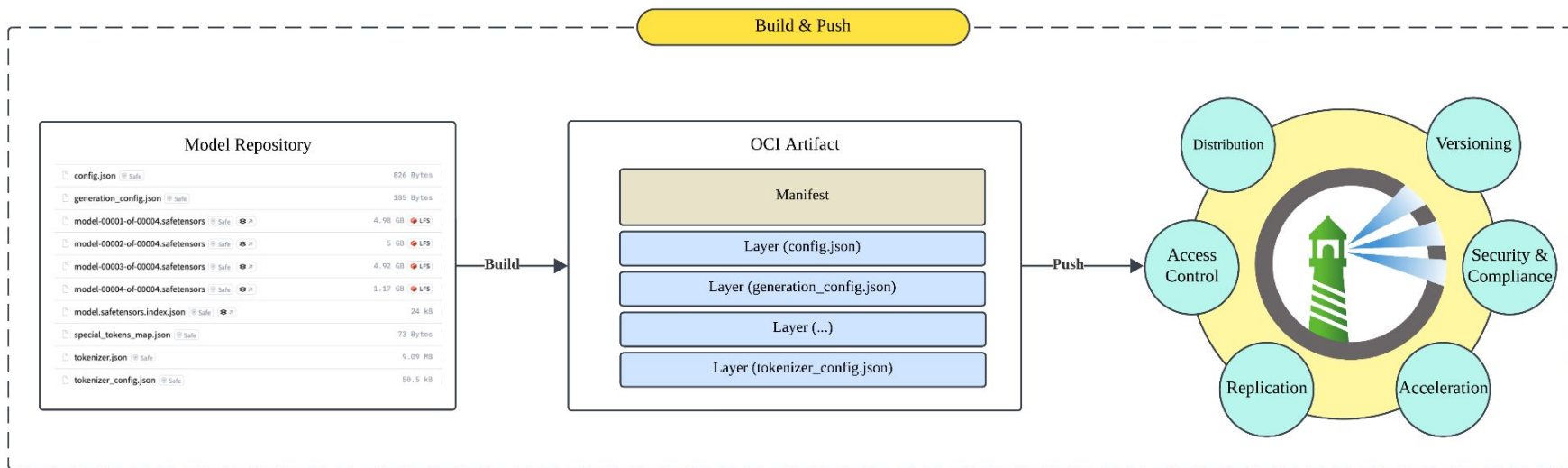
```
{
  "schemaVersion": 2,
  "config": {
    "mediaType": "application/vnd.caicloud.model.config.v1alpha1+json",
    "digest": "sha256:be948daf0e2f264ea70b713ea8db35850ae659c185706aa2fad74834455fe8c",
    "size": 187,
    "annotations": {
      "io.goharbor.artifact.v1alpha1.skip-list": "metrics,git",
      "io.goharbor.artifact.v1alpha1.version": "0.1.2",
      "io.goharbor.artifact.v1alpha1.name": "ML-Model",
      "io.goharbor.artifact.v1alpha1.xxx": "More metadata"
    }
  },
  "layers": [
    {
      "mediaType": "image/png",
      "digest": "sha256:d923b93eadd8af5c639a972710a4d919866aba5d8dfb4b9385099f70272da0",
      "size": 166015,
      "annotations": {
        "io.goharbor.artifact.v1alpha1.icon": ""
      }
    },
    {
      "mediaType": "<user_defined.mediaType>",
      "digest": "sha256:d923b93eadd8af5c639a972710a4d919866aba5d8dfb4b9385099f70272da0",
      "size": 166015
    }
  ]
}
```



# Overall idea: Build & Push

Wrap the AI models in OCI format and push to the OCI compliant registry.

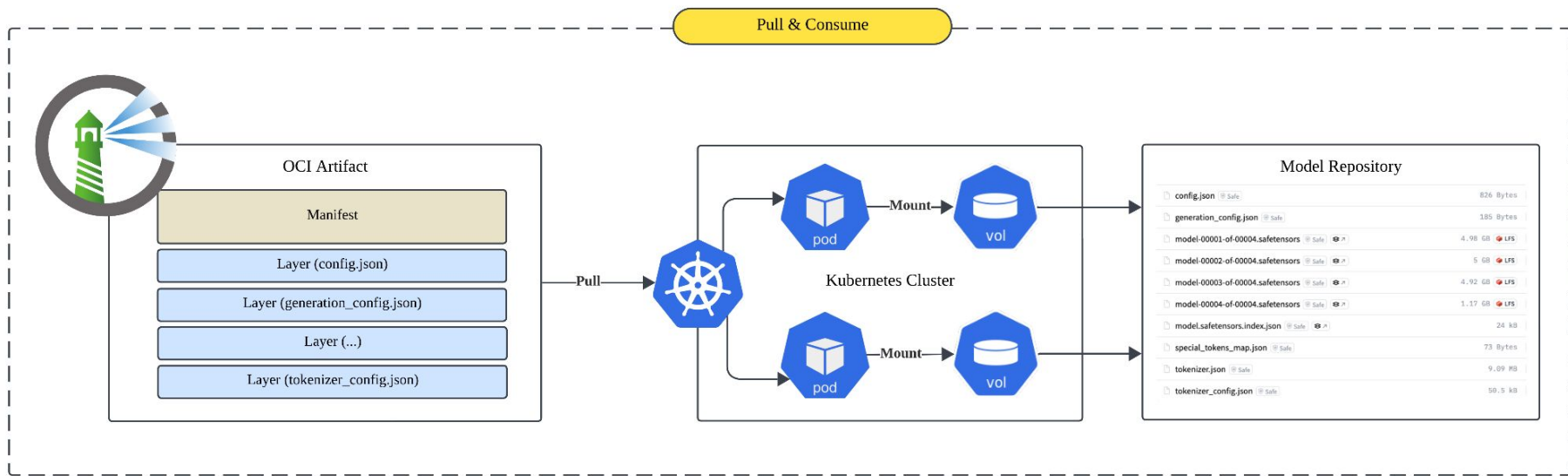
## Build & Push



# Overall Idea: Pull & Consume

Pull from OCI compliant registry and consume the AI models via image volume

## Pull & Consume



# Overall Idea: AI Model Spec

Materialize the AI model specification based on the extensibility of OCI artifact specification.

```
{
  "schemaVersion": 2,
  "mediaType": "application/vnd.oci.image.manifest.v1+json",
  "artifactType": "application/vnd.cncf.cnai.model.manifest.v1+json",
  "annotations": {
    "org.cnai.model.name": "my-model",
    "org.cnai.model.version": "0.1.0",
    "org.cnai.model.architecture": "transformer",
  }
  "config": {
    "mediaType": "application/vnd.oci.image.config.v1+json",
    "digest": "sha256:abcd1234ef56789...<truncated>",
    "size": 123
  },
  "layers": [
    {
      "mediaType": "application/vnd.oci.image.layer.v1.tar",
      "artifactType": "application/vnd.cncf.cnai.model.layer.v1.tar",
      "digest": "sha256:layer1digest...<truncated>",
      "size": 23456789,
      "annotations": {
        "org.cnai.model.config": "true"
      }
    },
    {
      "mediaType": "application/vnd.oci.image.layer.v1.tar",
      "artifactType": "application/vnd.cncf.cnai.model.layer.v1.tar",
      "digest": "sha256:layer2digest...<truncated>",
      "size": 34567890,
      "annotations": {
        "org.cnai.model.model": "true"
      }
    }
  ]
}
```

## More

- org.cnai.model.architecture
- org.cnai.model.family
- org.cnai.model.name
- org.cnai.model.format
- org.cnai.model.param.size
- ...

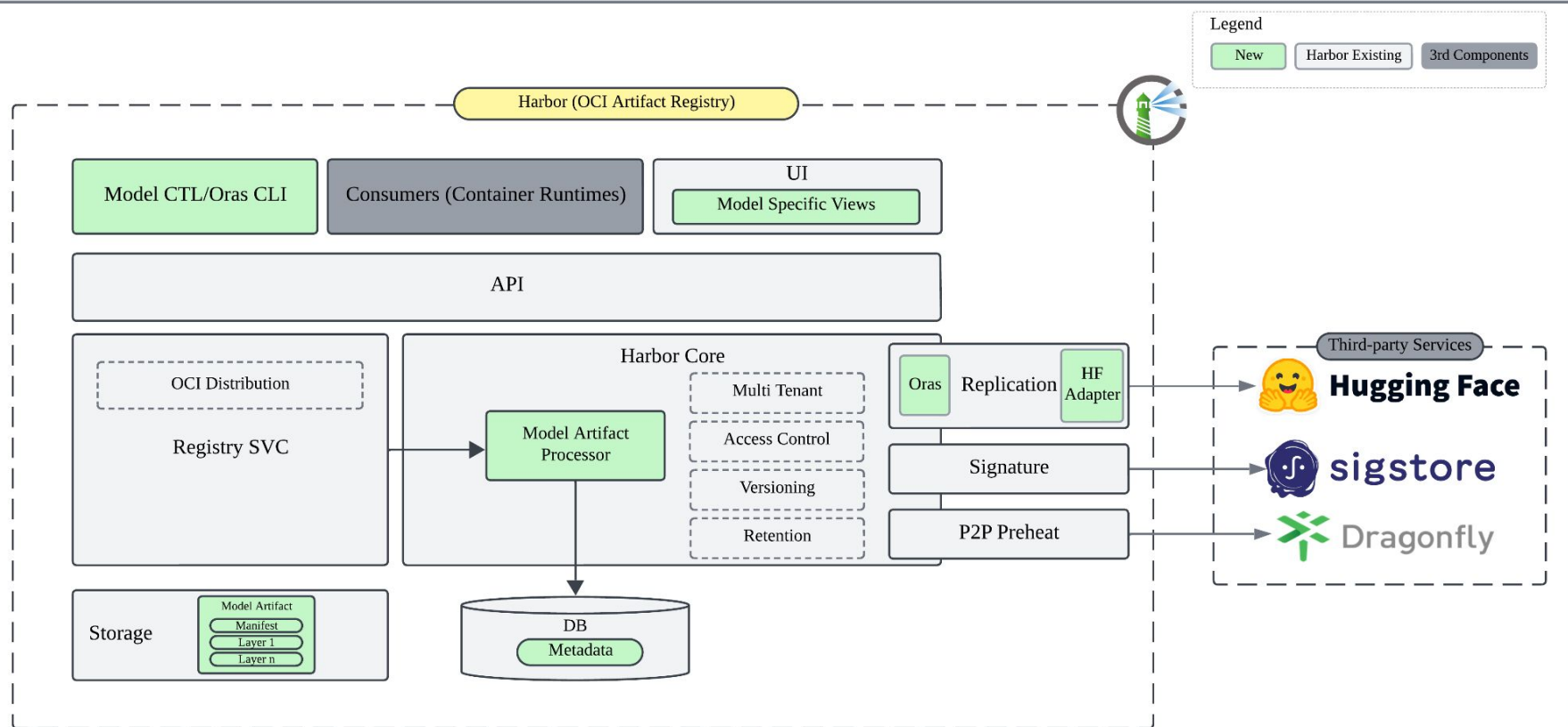
## Model files



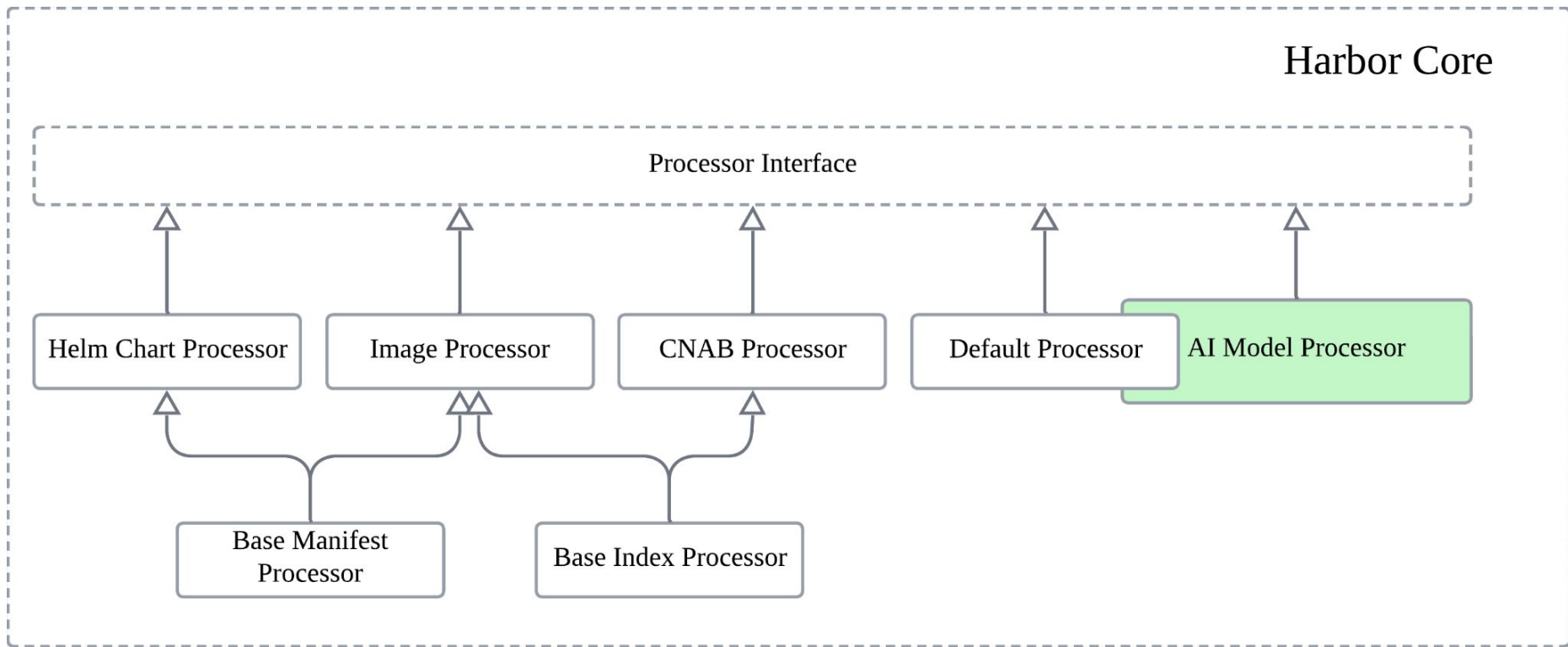
Welcome to discussions:  
[#model-spec-discussion](#)

# Implementation -HLD

## Overview



# Implementation - Backend



## Implementation - UI

Harbor
Search Harbor... English Default admin

# on - UI

Products < library < demo

## Name & Reference

sha256:56655434/Llama-2-7b-chat-hf

Authors NousResearch
Family llama3
Transformers
Pytorch
Safetensors
Llama

Text-generation
Facebook
Meta
Llama-2
En
Autotrain\_compatible

Text-generation-inference
Regionus

## Labels

Description

Meta developed and publicly released the Llama 2 family of large language models (LLMs), a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called Llama-2-Chat, are optimized for dialogue use cases. Llama-2-Chat models outperform open-source chat models on most benchmarks we tested, and in our human evaluations for helpfulness and safety, are on par with some popular closed-source models like ChatGPT and PaLM.

## Versioning Tags

Tags

COPY PULL COMMAND ▾

| <input type="checkbox"/>            | Name              | Pull Time        | Push Time        |
|-------------------------------------|-------------------|------------------|------------------|
| <input checked="" type="checkbox"/> | hugging-face-demo | 11/7/24, 3:51 PM | 11/6/24, 1:50 PM |

Page size 15 1 of 1 items

## Extra Metadata

**Overview**

|                                  |  |
|----------------------------------|--|
| org.cnsl.model.created           | 2023-07-18T19:45:53.000Z                               |
| org.cnsl.model.revision          | 351844e75ed0cbbe3f067fb3c808d2b8394ee                  |
| org.cnsl.model.title             | NousResearch/Llama-2-7b-chat-hf                        |
| org.cnsl.model.uri               | https://huggingface.co/NousResearch/Llama-2-7b-chat-hf |
| org.opencontainers.image.created | 2024-11-06T05:50:17Z                                   |

Additions

Files and versions

Model card

Files

- ☐ gitattributes
- ☐ LICENSE.txt
- ☐ README.md
- ☐ USE\_POLICY.md
- ☐ added\_tokens.json
- ☐ config.json
- ☐ generation\_config.json
- ☐ model-00001-of-00002.safetensors
- ☐ model-00002-of-00002.safetensors
- ☐ model.safetensors.index.json
- ☐ pytorch\_model-00001-of-00003.bin
- ☐ pytorch\_model-00002-of-00003.bin
- ☐ pytorch\_model-00003-of-00003.bin
- ☐ pytorch\_model\_bin.index.json
- ☐ special\_tokens\_map.json
- ☐ tokenizer.json
- ☐ tokenizer.model

## File List

\* LIGHT
Harbor API V2.0

Harbor
Search Harbor...
English ▾ Default ▾ admin ▾

---

**Projects**

- Logs
- Administration ▾
  - Users
  - Robot Accounts
  - Registries
  - Replications
  - Distributions
  - Labels
  - Project Quotas
  - Interrogation Services
  - Clean Up
  - Job Service Dashboard
  - Configuration

< Projects < library < demo

## sha256:56655434/Llama-2-7b-chat-hf

Author: NousResearch
Family: llama3
Transformers
Pytorch
Safetensors
Llama

Text generation
Facebook
Meta
Llama-2
En
Autotrain compatible

Text-generation-inference
Regions

### Description

Meta developed and publicly released the Llama 2 family of large language models (LLMs), a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called Llama-2-Chat, are optimized for dialogue use cases. Llama-2-Chat models outperform open-source chat models on most benchmarks we tested, and in our human evaluations for helpfulness and safety, are on par with some popular closed-source models like ChatGPT and PaLM.

### Tags

+ ADD TAG
REMOVE TAG

COPY PULL COMMAND ▾

| <input type="checkbox"/>            | Name              | Pull Time        | Push Time        |
|-------------------------------------|-------------------|------------------|------------------|
| <input checked="" type="checkbox"/> | hugging-face-demo | 11/7/24, 3:51 PM | 11/6/24, 1:50 PM |

Page size 15 1 of 1 items

### Overview

| Overview                         |  |
|----------------------------------|--|
| org.chnai.model.created          | 2023-07-18T19:45:53.000Z                               |
| org.chnai.model.revision         | 35f844e75ed0bcbbef067fb3c808d2b3894ee                  |
| org.chnai.model.title            | NousResearch/Llama-2-7b-chat-hf                        |
| org.chnai.model.uri              | https://huggingface.co/NousResearch/Llama-2-7b-chat-hf |
| org.opencontainers.image.created | 2024-11-06T05:50:17Z                                   |

### Additions

**Model card** Files and versions

# Readme

## Llama 2

Llama 2 is a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. This is the repository for the 7B fine-tuned model, optimized for dialogue use cases and converted for the Hugging Face Transformers format. Links to other models can be found in the index at the bottom.

### Model Details

*Note: Use of this model is governed by the Meta license. In order to download the model weights and tokenizer, please visit the website and accept our License before requesting access here.*

Meta developed and publicly released the Llama 2 family of large language models (LLMs), a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. Our fine-tuned LLMs, called Llama-2-Chat, are optimized for dialogue use cases. Llama-2-Chat models outperform open-source chat models on most benchmarks we tested, and in our human evaluations for helpfulness and safety, are on par with some popular closed-source models like ChatGPT and PaLM.

#### Model Developers Meta

Variations Llama 2 comes in a range of parameter sizes — 7B, 13B, and 70B — as well as pretrained and fine-tuned variations.

Input Models input text only.

Output Models generate text only.

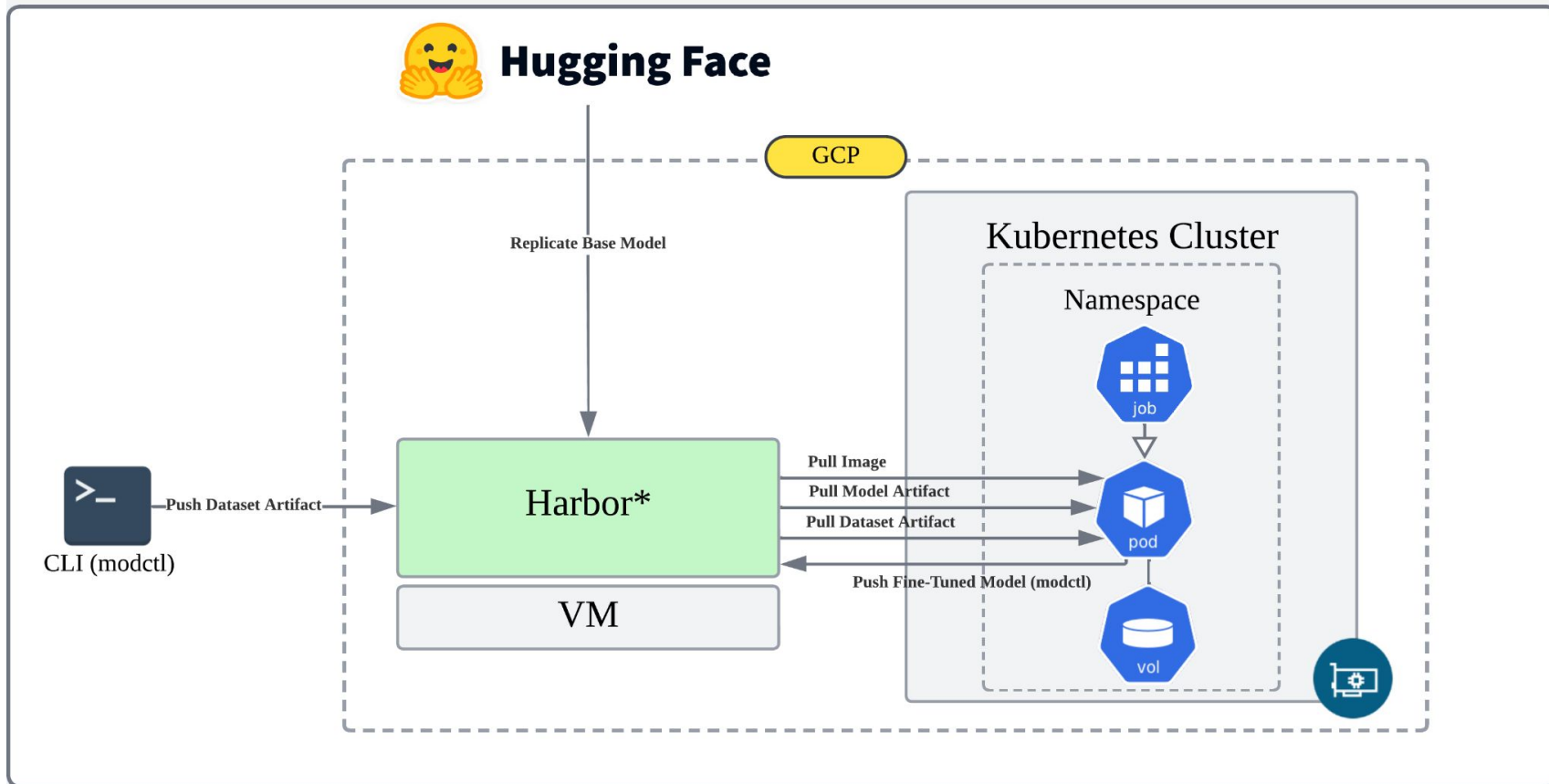
\* LIGHT

Harbor API V2.0

A set of features that are applicable to AI models

- OCI Compatible (Support image volume)
- Multi Tenant
- Versioning
- Access Control (RBAC, Policy Control)
- Signature
- Replication
  - harbor<-> harbor
  - harbor<-HF
- Retention
- P2P Preheat
- GC
- Audit log
- Quota

## Demo Env/Topology







## Security and Compliance

---

Integrate model scanner through pluggable scanner framework to support vulnerability scanning, license checking etc.



## Dependency Management

---

Set up linkages between models and other related resources (DS, Images) with artifact accessories mechanism.



## Search and Discovery

---

Enhance API to allow for cataloging and tagging to making it easier to discover and organize models across teams.



## Replication +

---

Implement bidirectional replication with upstream model registries to support publishing the ready models.



**KubeCon**



**CloudNativeCon**

**North America 2024**

**THANK YOU!**

