
KubeCon
CloudNativeCon
North America 2024

KubeCon

CloudNativeCon

North America 2024

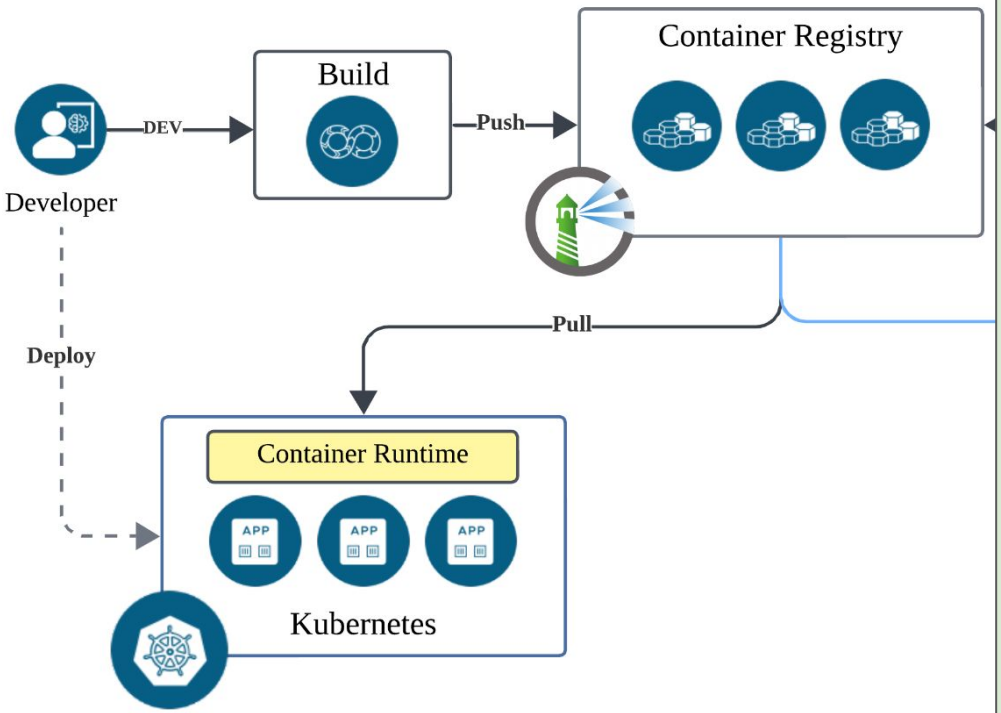
Managing and Distributing AI Models Using OCI Standards and Harbor

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



Kubernetes: Platform for container orchestration and workload management

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

AI Workload: code + model, dataset

Motivations

Private AI model registry service is a hard requirement.

Intelligent Property Protection

Security & Compliance

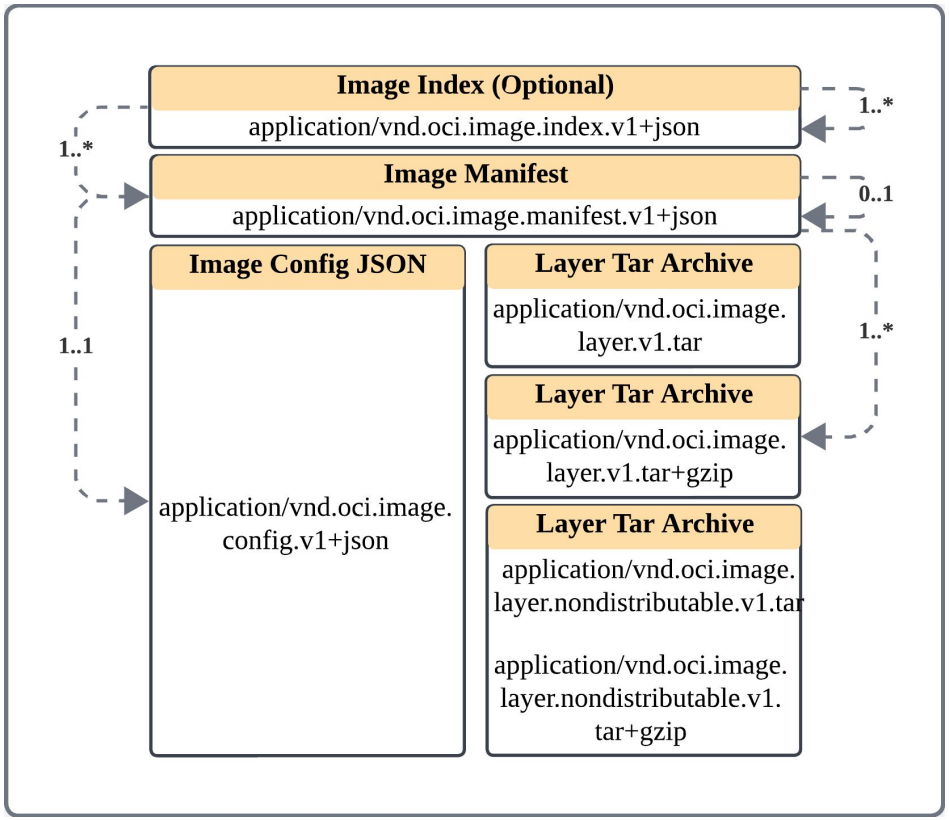
Performance

Cost Efficiency

Smooth Integration

Background - OCI Spec

OCI Image Spec



Characteristics:

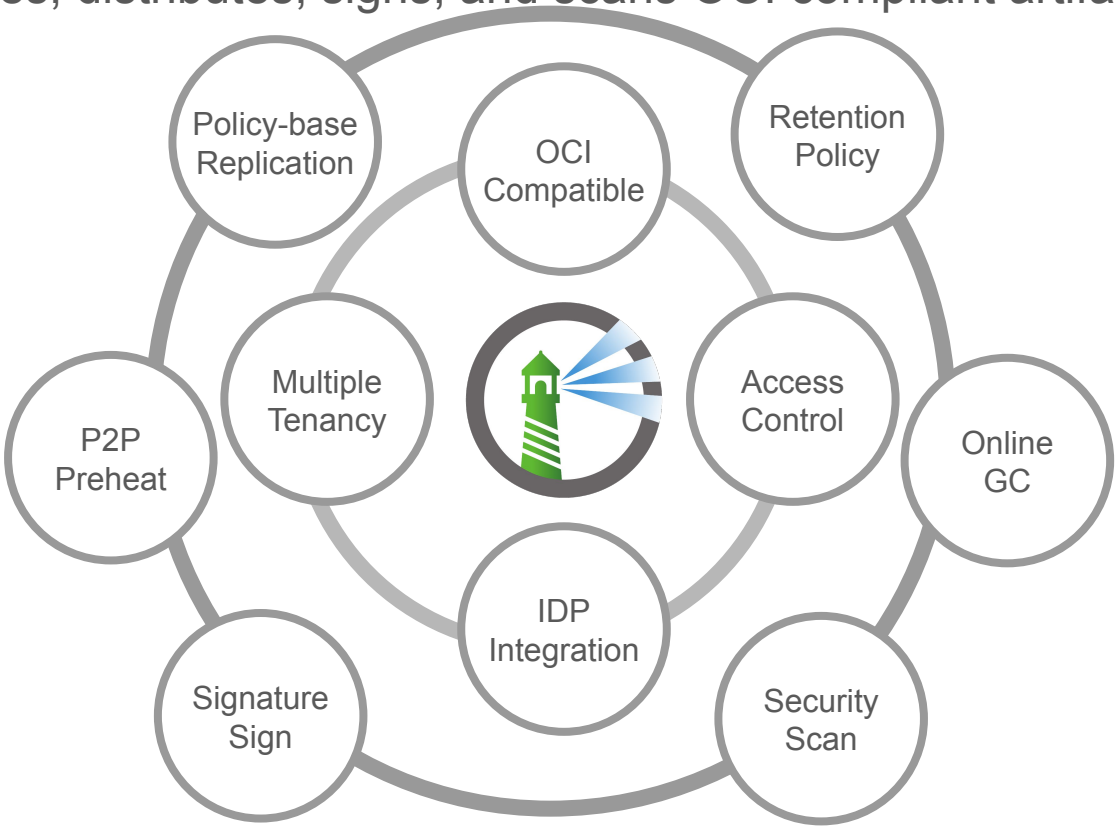
- Standardized Format
- Layered Structure
- Metadata Manifest
- Compatibility

Benifits:

- Reusability of Infrastructure
- Compatibility with Existing Ecosystems
- Portability

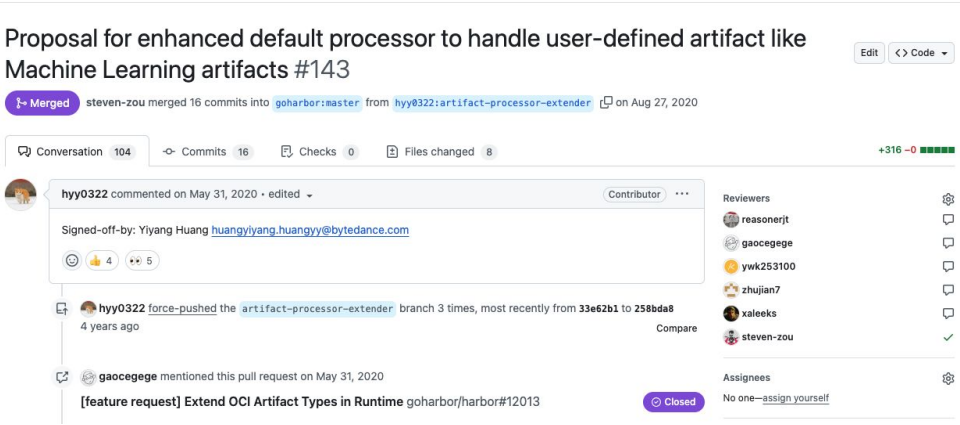
Background - Harbor

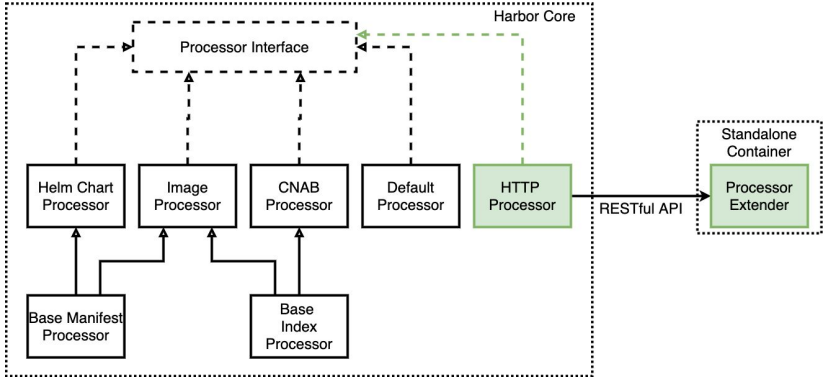
Harbor, a CNCF-graduated, open-source trusted cloud native registry project; stores, distributes, signs, and scans OCI compliant artifacts.



History Related to AI/ML Model

Harbor has a [long history](#) to start the AI model Jurney.

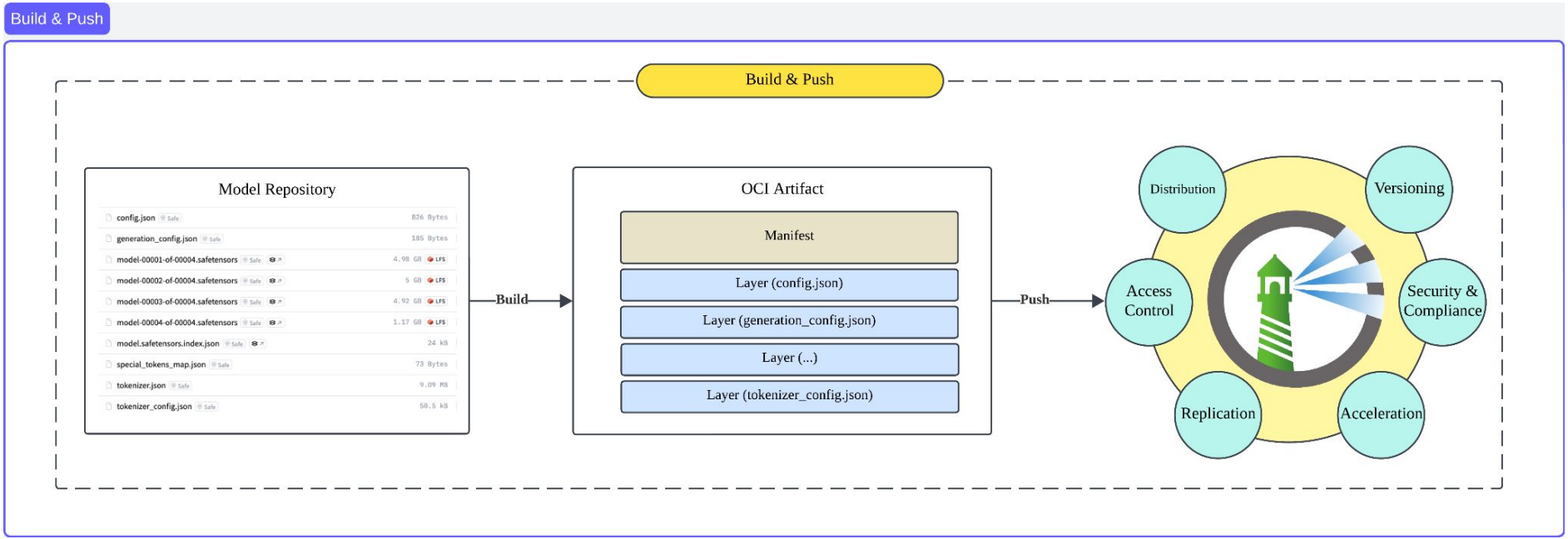




Collaborate with Harbor community partner CaiCloud (aka. ByteDance Volcano Engine)

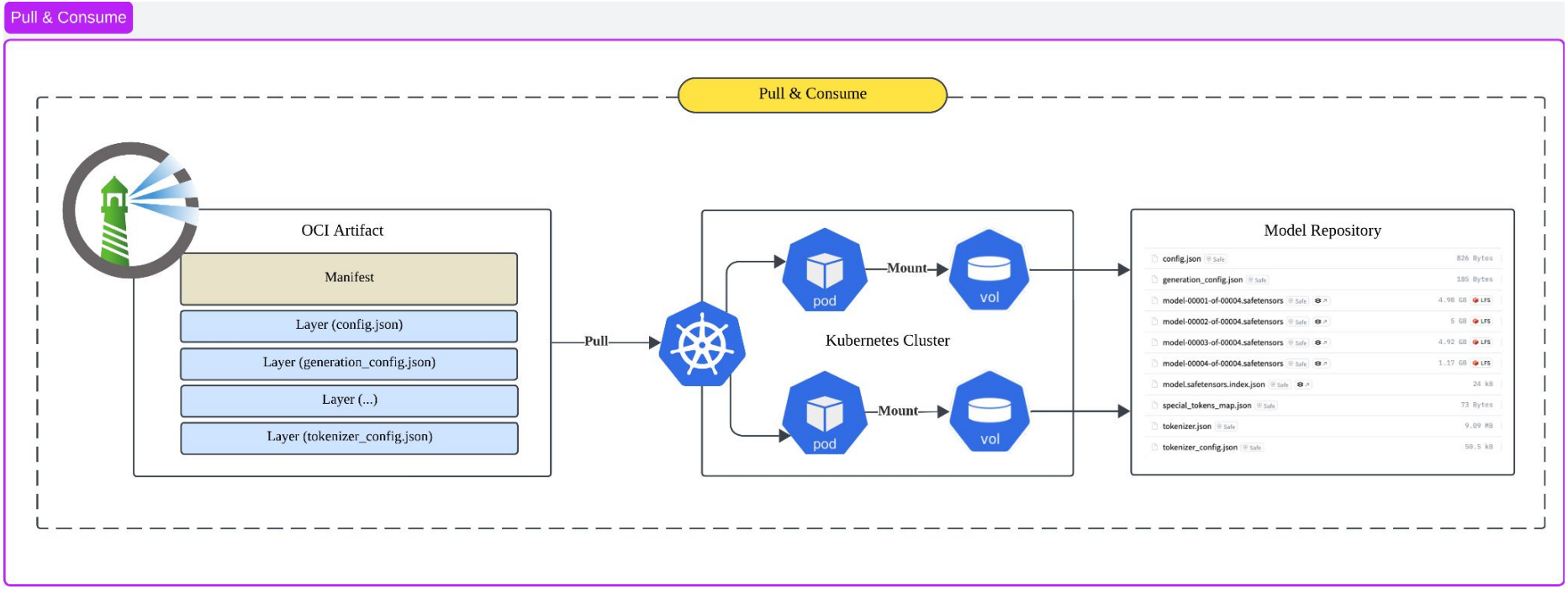
Overall idea: Build & Push

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



Overall Idea: Pull & Consume

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



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

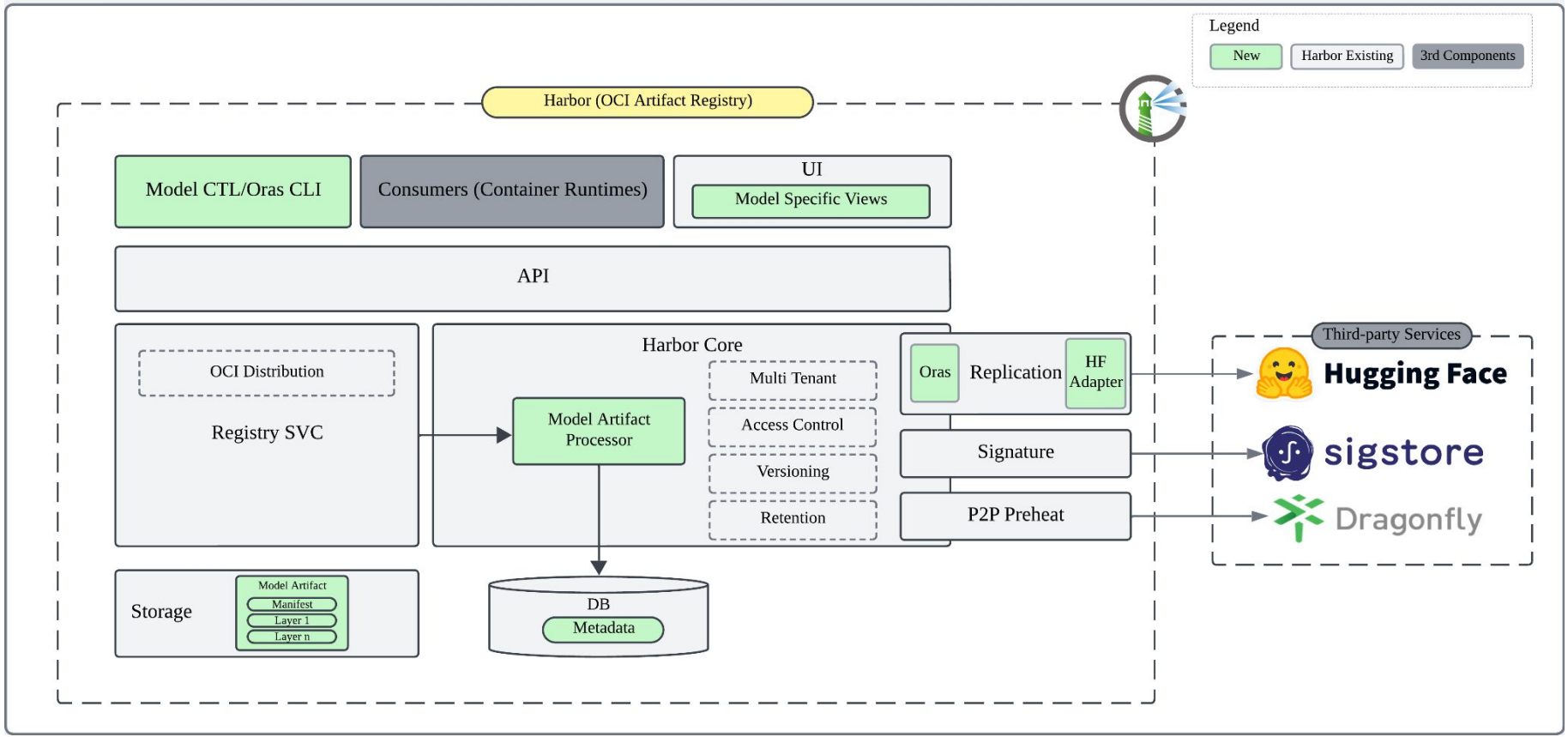
- org.cnai.model.architecture
- org.cnai.model.family
- org.cnai.model.name
- org.cnai.model.format
- org.cnai.model.param.size
- …

Model files

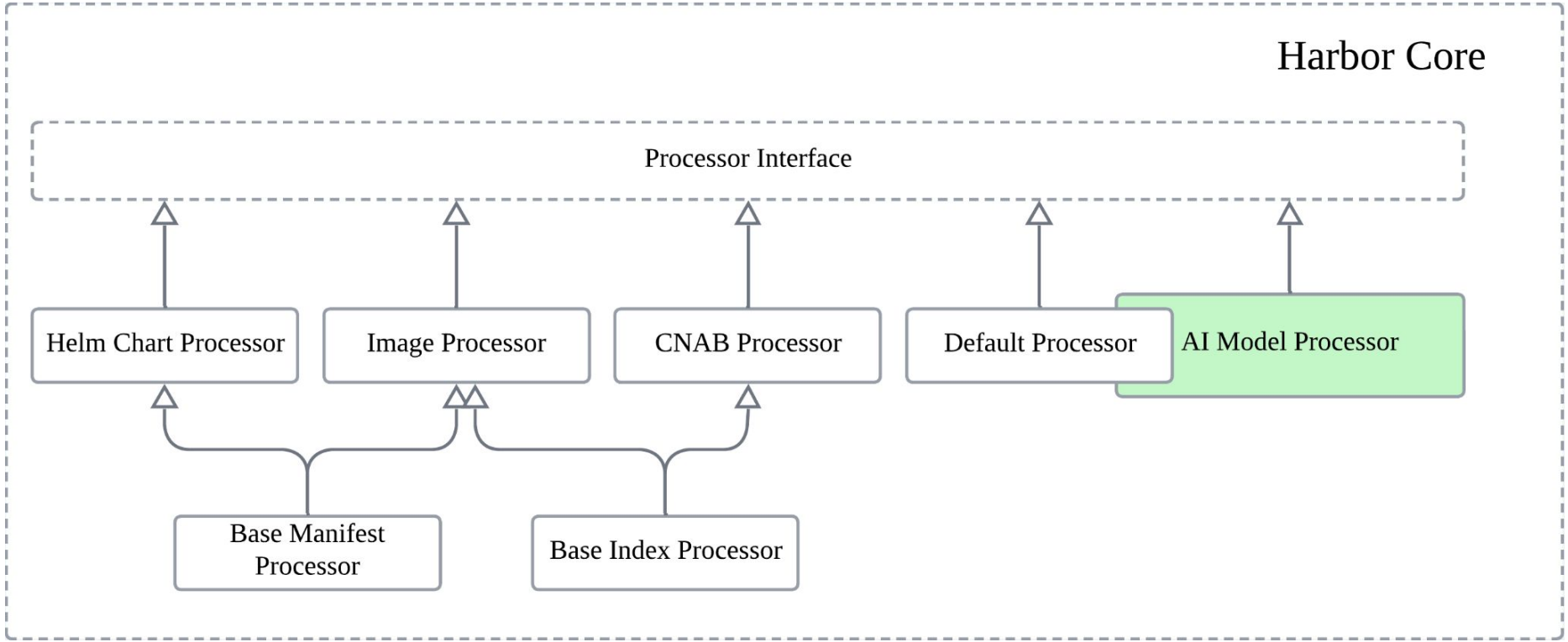
Welcome to discussions: #model-spec-discussion

Implementation -HLD

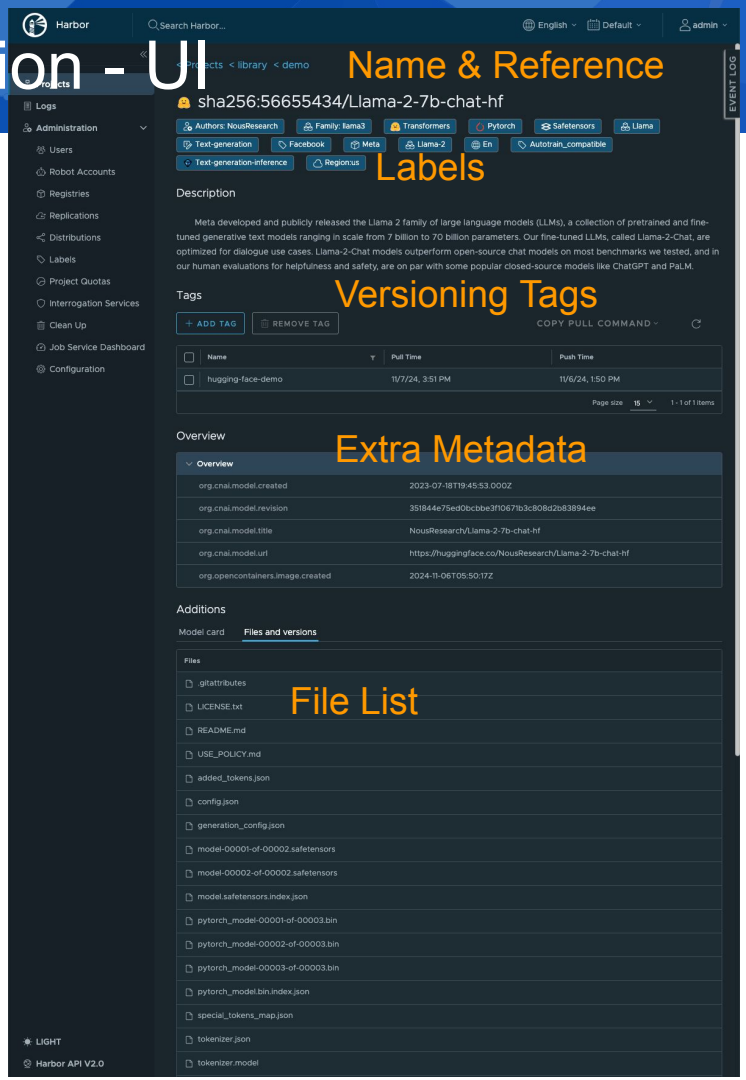
Overview

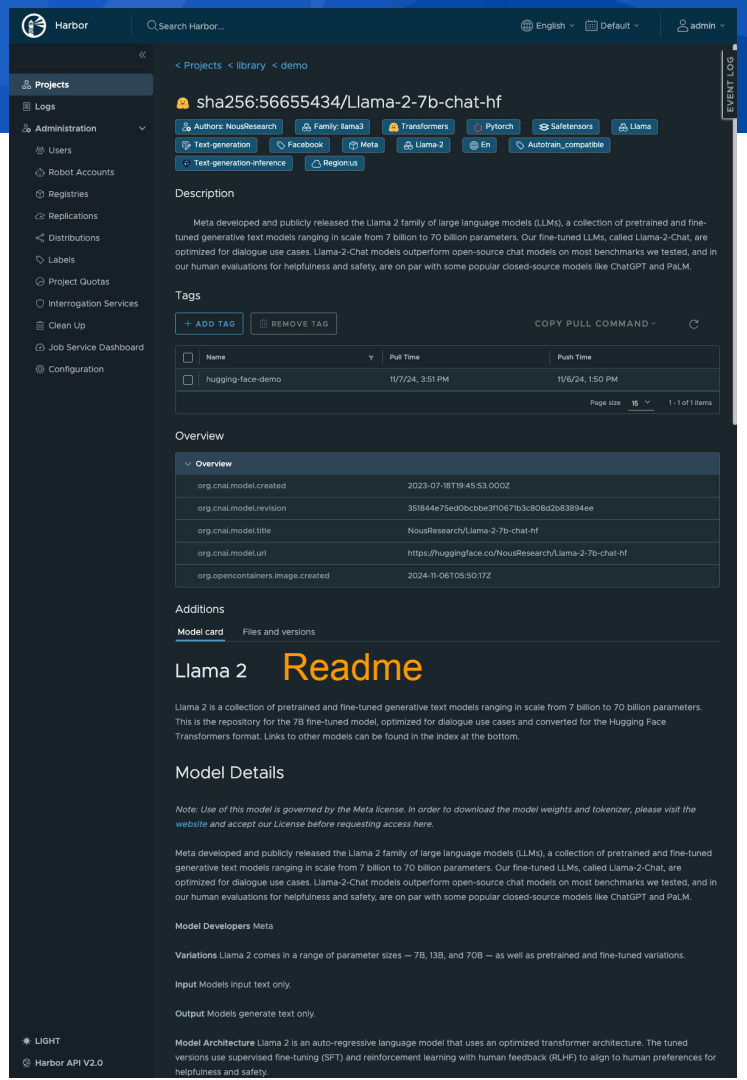


Implementation - Backend



Implementation - UI





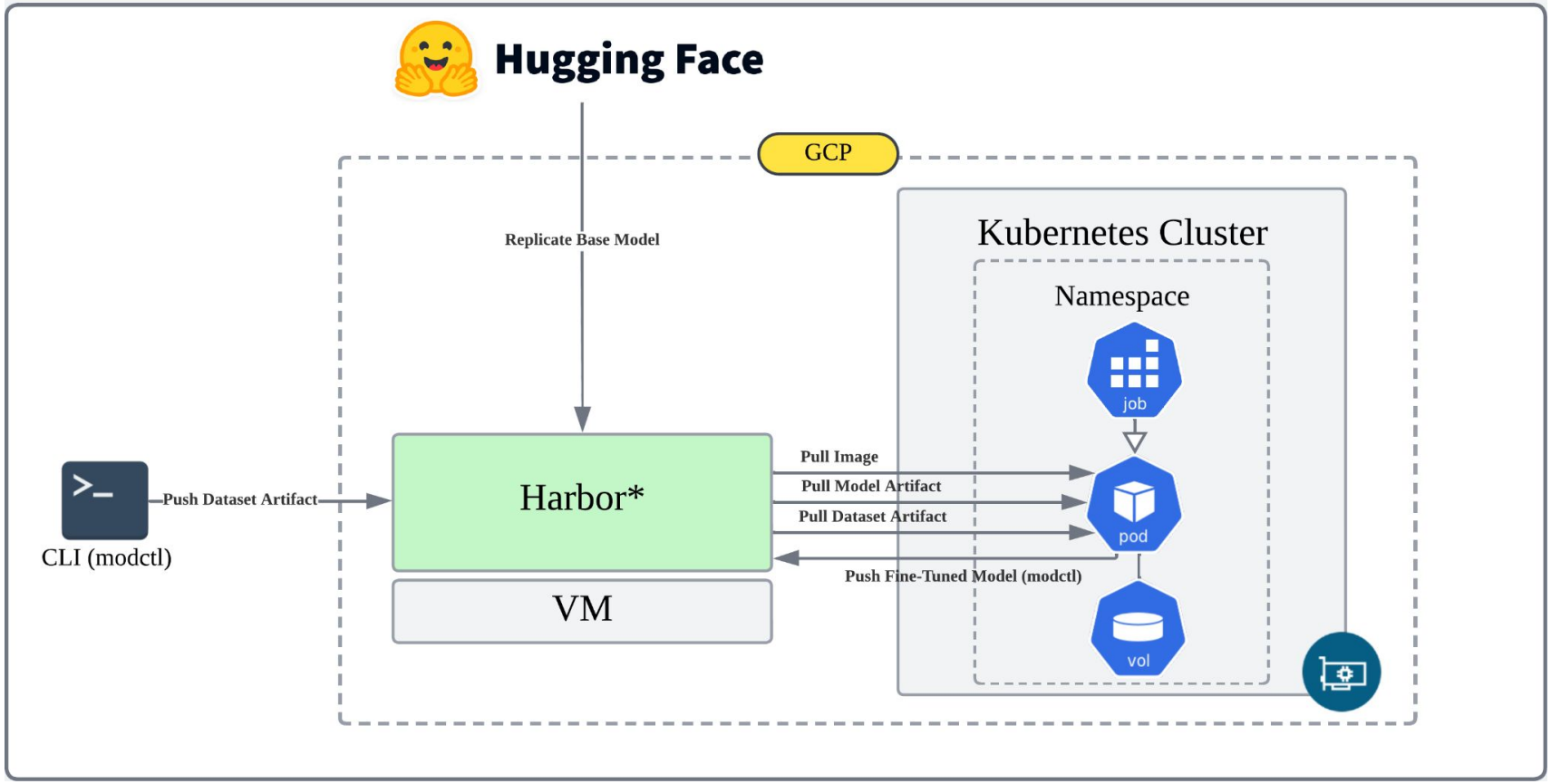
Features

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 Time

Demo Env/Topology



Next

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!