KubeCon CloudNativeCon India 2026

#KubeCon #CloudNativeCon

ModelPack: Bringing the OCI Standard to AI/ML

Andrew Block
Distinguished Architect, Red Hat
ModelPack Maintainer

AI Model Management Challenges

Concerns throughout various phases in the lifecycle of a model

Model Packaging

- Various types of content
  - Files, folders, archives
- Container images used as a vehicle containing model content
- Models stored with proprietary metadata locking them into specific ecosystems

Model Storage

- Models are stored in a variety of locations
  - Git
  - S3
  - Proprietary Model Stores

Model Serving

- Framework specific implementations results in unique architectures for each model
- Python based microservice wrappers are hard to scale along with manage concurrency effectively
- Kubernetes based abstractions increases complexity while demanding additional compute resources

ModelPack

Vendor neutral, **Open Standard** for managing **AI/ML Models**

CLOUD NATIVE COMPUTING FOUNDATION

Sandbox project since May 2025

Leveraging the Power of OCI Artifacts

ModelPack packages AI/ML Models as **OCI Artifacts** using a standardized and defined specification for easy **distribution** and **consumption**

Benefits

Apply **Cloud Native** standards and patterns

Reuse existing infrastructure

Interoperability with existing tooling and frameworks

OCI Artifacts enables the storage of arbitrary content using the same framework and tooling as traditional container images

Features and Capabilities

Standard Packaging

Uniform method for assembling model files, dependencies, and metadata

CLI Enabled

`modctl` CLI enables effective management of AI content

CI/CD Integration

Easily integrate with existing tools and work

Security Built in

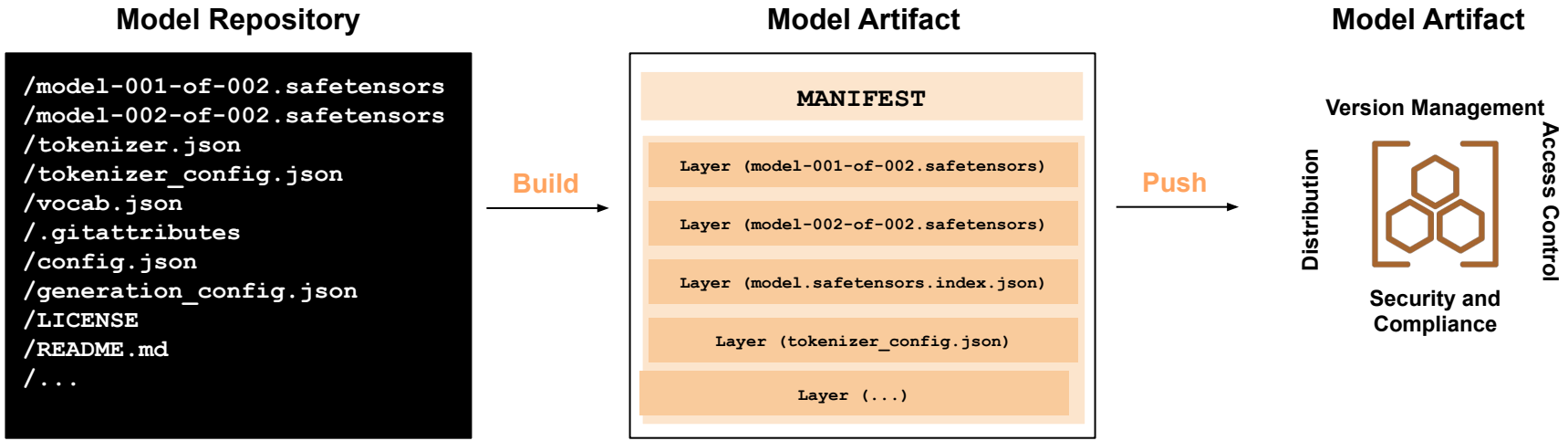
Distributing as OCI artifacts enables the reuse existing tooling and approaches (eg: signatures, SBOM's)

Community First

Built in the community, for the community to standardize and accelerate the use of AI content

Building and Publishing Workflow

Utilize build tools to assemble required resources into an **OCI artifact** aligning to the *model format specification*. Generated artifact can be published to a **OCI registry**, such as Docker Hub or Quay



Retrieval and Runtime

Several options are available to **retrieve** and **run** a ModelPack formatted **artifact**

Command Line Tools

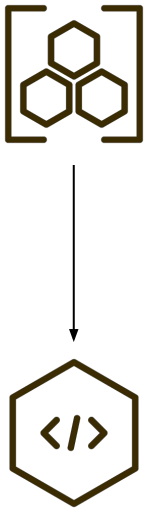
`modctl` and other Cloud Native tools can be used to inspect and retrieve ModelPack content

Container Runtime

ModelPack content retrieved, stored, and managed using native container runtimes

Container Storage Interface (CSI)

Support to natively integrate into the storage framework for runtime in Kubernetes



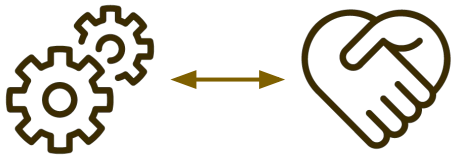
A Focus on Integrations

Docker Model Runner (DMR)

First class support for **producing** and **running** ModelPack formatted models

CNCF Inner-Loop for AI Engineers Initiative

ModelPack participating as a core component with the goal of enabling AI engineering efforts locally



Partnering with **projects**, **organizations** and the **community**

Let us know if you want to integrate collaborate!

Resources

Opportunities to **Collaborate** with the ModelPack **Community**

Web

modelpack.org

GitHub

github.com/modelpack

Slack

#modelpack

Getting Involved

Connecting with the **Community** Has Never Been Easier!

Collaborate on GitHub

Participate in the Community Meetings

Join the Project Slack Channel


KubeCon
CloudNativeCon
India 2026