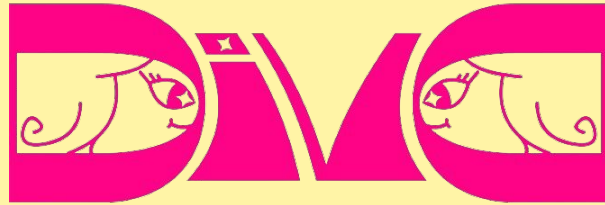


Kubernetes and AI

Ceyda Duzgec



Diva AI 2024



Backend Developer



Ceyda Düzgeç

About me

- B.Sc. Computer Science & Engineering @Sabancı
Final project => 🦴 Medical Image Classification
- M.Sc. Software Engineering @Bogazici
- Backend Developer -> Cloud & Platform Engineer
📚 AWS and Kubernetes



Cloud & Platform
Engineer

<https://www.sufle.io>

sufle



Agenda

01 Kubernetes

02 Kubectl-ai

03 K8sgpt

04 Kubeflow

05 Local AI

slides





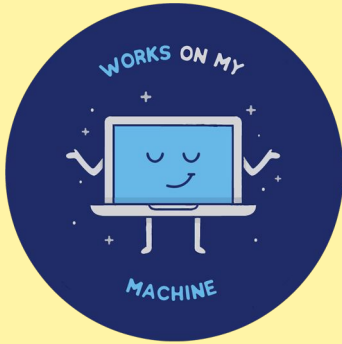
01

Kubernetes

slides



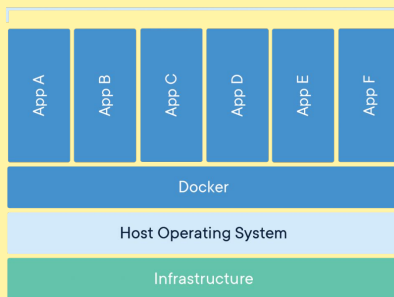
Let's start with containers



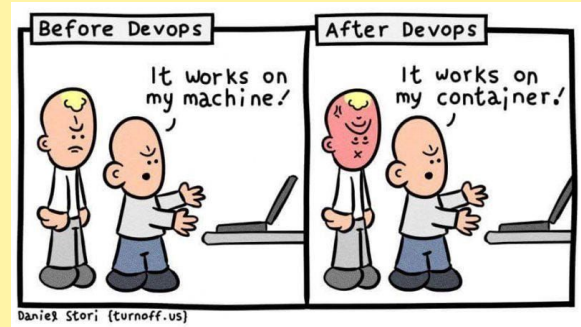
Role in AI workflows:

- Reproducibility
- Dependency management

Containerized Applications



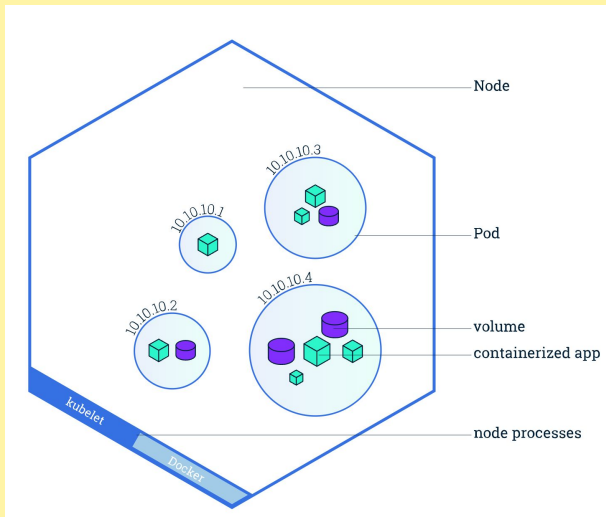
- Lightweight, portable and consistent environments
- Isolation and resource efficiency





What is Kubernetes?

Kubernetes, also known as **K8s** or **Kube**, is an open source system for automating deployment, scaling, and management of containerized applications.

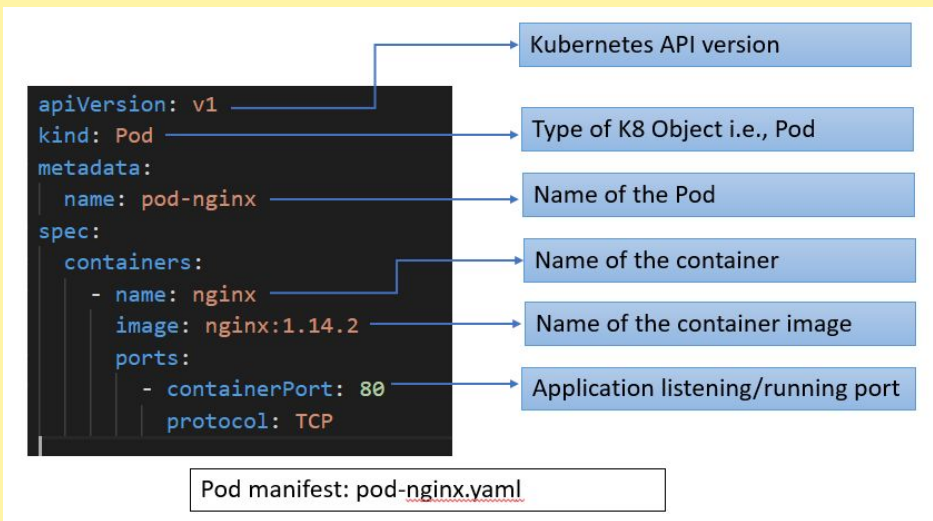


- **Pods:** The smallest deployable units in Kubernetes, consisting of one or more containers.
- **Nodes:** Worker machines that run containerized applications.
- **Deployment:** A Kubernetes resource that ensures a specified number of identical Pods are running, automatically managing updates and rollbacks with zero downtime using rolling updates.

<https://kubernetes.io/docs/tutorials/kubernetes-basics/explore/explore-intro/>

Manifest ✨ file

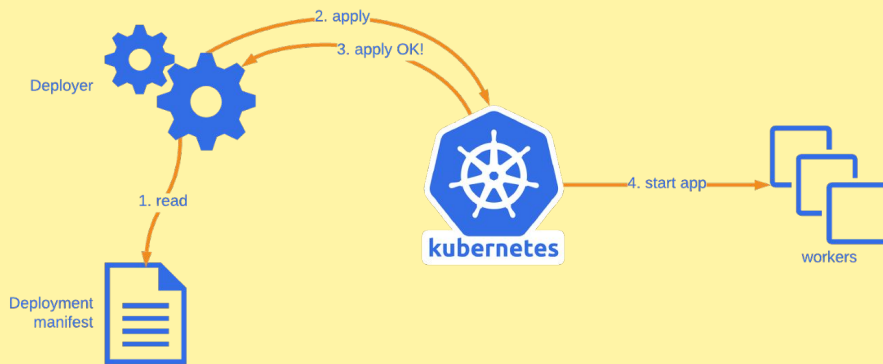
Manifest File: A Kubernetes manifest file is a YAML (or JSON) configuration file used to define, manage and describe the desired state of a Kubernetes resources. This includes details like resource type, metadata, and specifications.



Manifest file

Declarative Management: Kubernetes operates on a declarative model, meaning you declare the desired state of your resources in the manifest files, and Kubernetes works to maintain that state.

Automation: Facilitates the automation of deployment, scaling, and management of containerized applications.



Benefits of K8s

- Scalability
- Automated Deployment and Updates
- Resource Management
- Portability
- High Availability
- Monitoring and Logging
- Easy Integration
- Cost Optimization

Case Study

OpenAI began running Kubernetes on top of AWS

2016

"Building the Infrastructure that Powers the Future of AI" presented by Vicki Cheung, Member of Technical Staff & Jonas Schneider, Member of Technical Staff at OpenAI from KubeCon/CloudNativeCon Europe 2017



<https://kubernetes.io/case-studies/openai/>

2017

We've been running Kubernetes for deep learning research for over two years. While our largest-scale workloads manage bare cloud VMs directly, Kubernetes provides a fast iteration cycle, reasonable scalability, and a lack of boilerplate which makes it ideal for most of our experiments.

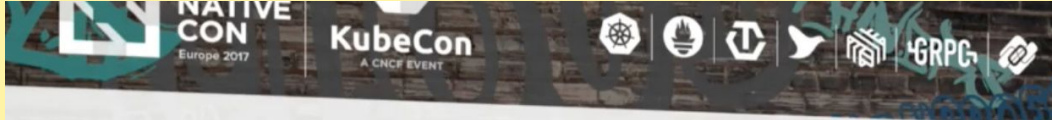
2018

<https://openai.com/index/scaling-kubernetes-to-2500-nodes/>

2021

We've found Kubernetes to be an exceptionally flexible platform for our research needs. It has the ability to scale up to meet the most demanding workloads we've put on it. There are many areas yet though where it needs improvement, and the Supercomputing team at OpenAI will continue to explore how Kubernetes can scale.

<https://openai.com/index/scaling-kubernetes-to-7500-nodes/>



So, Why Kube?

- Kube: core set of interfaces & abstractions for infrastructure
- Provides flexibility for user workloads and infra operations
- We're excited for the future:
 - **scheduling**: GPUs, fair scheduling, priorities
 - **introspection**: better visibility & monitoring
 - **scalability**: 10K nodes and beyond



02

Kubectl-ai

slides



Kubectl OpenAI Plugin

This project is a kubectl plugin to generate and apply Kubernetes manifests using OpenAI GPT.

Homebrew

Add to `brew` tap and install with:

```
brew tap sozercan/kubectl-ai https://github.com/sozercan/kubectl-ai  
brew install kubectl-ai
```

 14

 73

 996

```
export OPENAI_API_KEY=<your OpenAI key>  
export OPENAI_DEPLOYMENT_NAME=<your OpenAI deployment/model name, defaults to "gpt-3.5-turbo-0301">  
export OPENAI_ENDPOINT=<your OpenAI endpoint, like "https://my-aoi-endpoint.openai.azure.com" or "http://localhost:8080/v1">
```

<https://github.com/sozercan/kubectl-ai>

Kubectl OpenAI Plugin

```
$ kubectl ai "create an nginx deployment with 3 replicas" 
```

```
✦ Attempting to apply the following manifest:
```

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx-deployment
  labels:
    app: nginx
spec:
  replicas: 3
  selector:
    matchLabels:
      app: nginx
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
        - name: nginx
          image: nginx:1.7.9
          ports:
            - containerPort: 80
```

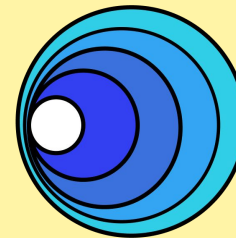
```
Use the arrow keys to navigate: ↓ ↑ → ←
```

```
? Would you like to apply this? [Reprompt/Apply/Don't Apply]:
```

```
+ Reprompt
  ▶ Apply
  Don't Apply
```

03

K8sgpt



slides





Kubernetes Superpowers



Giving Kubernetes Superpowers to everyone

👁️ 55

- Provides detailed and contextualized explanations of Kubernetes error messages in simple English
- Provides insights, diagnoses
- Supports different AI backends: OpenAI, Azure, Cohere, Amazon Bedrock, Google Gemini and local models.

🍴 593



★ 5.2k

```
k8sgpt auth list
Default:
> openai
Active:
> openai
Unused:
> localai
> azureopenai
> cohere
> amazonbedrock
> amazonsagemaker
> google
> noopai
> huggingface
> googlevertexai
> oci
```



<https://github.com/k8sgpt-ai/k8sgpt>

Some tools for AIOps

- HolmesGPT - The Open Source On-Call/DevOps Agent
=> <https://github.com/robusta-dev/holmesgpt>
- Kubernetes Copilot
=> <https://github.com/feiskyer/kube-copilot>
- Botkube
=> <https://github.com/kubeshop/botkube>
- Kubectl-GPT
=> <https://github.com/devinjeon/kubectl-gpt>
- Mico
=> <https://github.com/tahtaciburak/mico>
- KoPylot: Your AI-Powered Kubernetes Assistant 
=> <https://github.com/avsthiago/kopylot>
- Kopilot : Your AI Kubernetes Expert
=> <https://github.com/knight42/kopilot>

04

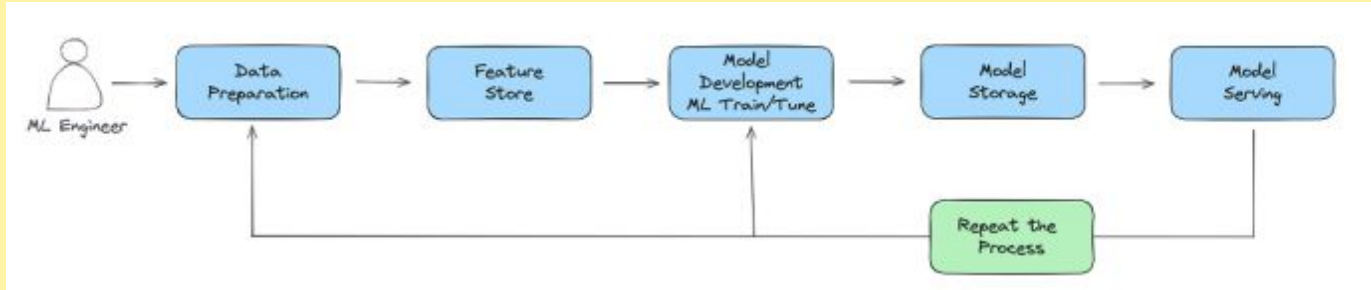
Kubeflow



slides



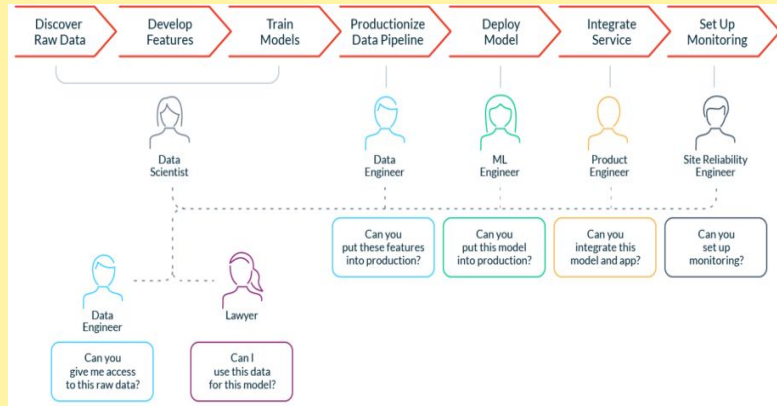
ML Pipeline



The typical ML pipeline is comprised of:

- Data Preparation (collection, cleaning/pre-processing, feature engineering)
- Model Training (model selection, architecture, hyperparameter tuning)
- CI/CD, Model Registry (storage)
- Model Serving
- Observability (usage load, model drift, security)

Challenges for ML



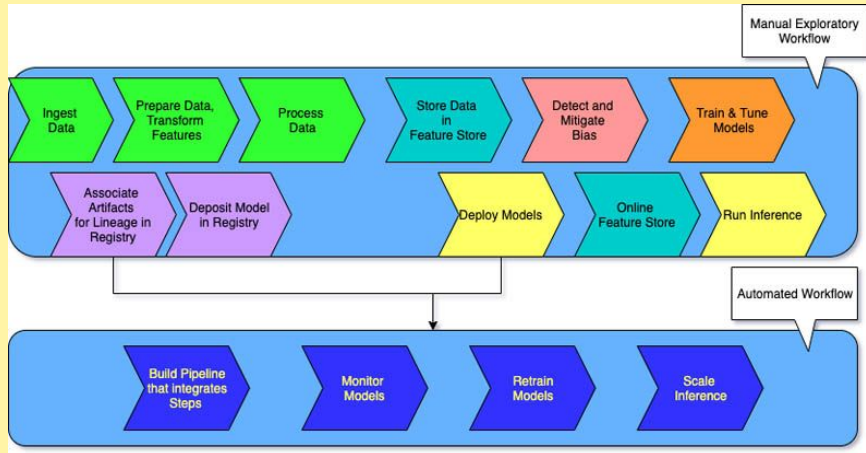
Data Preparation: Data Size, Data Synchronization, Data Governance

Model Training: Rising Processing Demands, Cost Efficiency, Scalability, Orchestration/Scheduling, Custom Dependencies

Model Serving: Microservice Architecture and Developer Experience, Resource Allocation

Cross-Cutting Concerns: Cost Control, Observability, Disaster Recovery and Business Continuity, Security and Compliance Audits

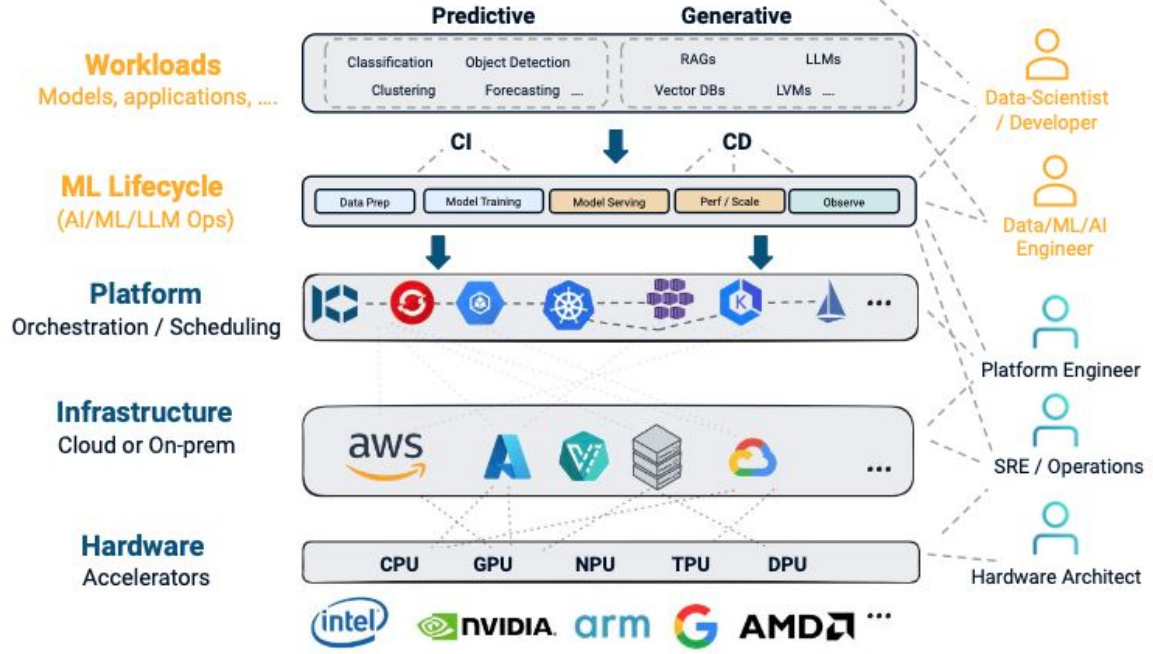
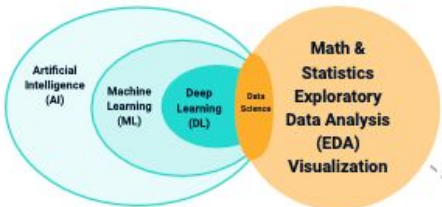
What is MLOps?



Machine learning operations (MLOps) are a set of practices that automate and simplify machine learning (ML) workflows and deployments

- **CI/CD:** Continuous integration and continuous deployment for rapid, reliable delivery of ML models.
- **Monitoring:** Track model performance and health in production.
- **Governance:** Ensure compliance and manage model versions.

Role of Kubernetes: Provides the underlying infrastructure to automate and scale MLOps pipelines.



Feature	Focused Tools (local/on-prem)	Managed Platforms	“Kube Native”
Examples	MLflow, Backyard AI, Ollama, Hugging Face TGI	AWS SageMaker, scale.ai	Kubeflow / KServer (Hugging Face TGI / LocalAI)
Scope	Specific functionalities within ML lifecycle	Managed MLOps service	End-to-end MLOps platform
Open Source	Yes	No	Yes
Scalability & Portability	Moderate	Depends on cloud provider	High
Setup & Management	Simpler	Simpler	Complex
Portability	Mostly Machine based	Mostly Cloud	Everywhere
Vendor Lock-in	No	Yes (to specific cloud provider)	No

Cloud Native Artificial Intelligence



[About](#) [Projects](#) [Training](#)

WHO WE ARE

The Cloud Native Computing Foundation (CNCF) hosts critical components of the global technology infrastructure.

We bring together the world's top developers, end users, and vendors and run the largest open source developer conferences. CNCF is part of the nonprofit [Linux Foundation](#).

Cloud Native Artificial Intelligence is an evolving extension of Cloud Native.

Cloud Native Artificial Intelligence (CNAI) refers to approaches and patterns for building and deploying AI applications and workloads using the principles of Cloud Native. Enabling repeatable and scalable AI-focused workflows allows AI practitioners to focus on their domain.

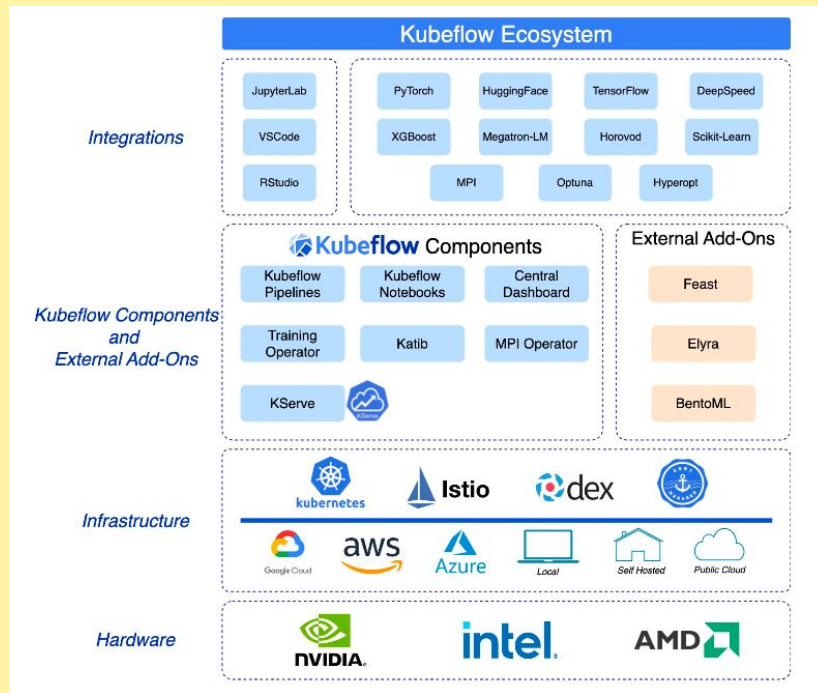
Orchestration - Kubeflow

Kubeflow is an example of a CNAI tool supporting ML Operations (MLOps). Using technologies such as Kubernetes, stateless architectures, and distributed systems, Kubeflow helps AI/ML communities adopt Cloud Native tools more efficiently. The successful adoption of Kubeflow highlights the successful integration of Cloud Native technologies for AI/ML/DL. Kubeflow has been highly progressive in its ability to apply machine learning concepts to elastic substrates provided by Kubernetes, with many other projects following suit.⁷² Kubeflow follows Kubernetes best practices and applies them to the AI/ML space, such as declarative APIs, composability, and portability. Kubeflow implements individual microservices for every stage of the ML lifecycle. For example, Kubeflow Training Operator is used for distributed training, Katib is used for hyperparameter tuning fine-tuning, and Kubeflow KServe is used for model serving. That allows users to integrate individual Kubeflow components into their ML infrastructure or use Kubeflow as an end-to-end ML platform.

<https://www.cncf.io/reports/cloud-native-artificial-intelligence-whitepaper/>

Why Kubeflow?

Kubeflow is an open-source project dedicated to making deployments of machine learning workflows on Kubernetes simple, portable, and scalable.




👁️ 361

🍴 2.3k

★ 13.9k

<https://github.com/kubeflow/kubeflow>

Fun fact

Swiss scientists have approved Kubeflow 

Kubeflow at CERN

Initial deployments and proof of concept in 2019

In production since April 2021

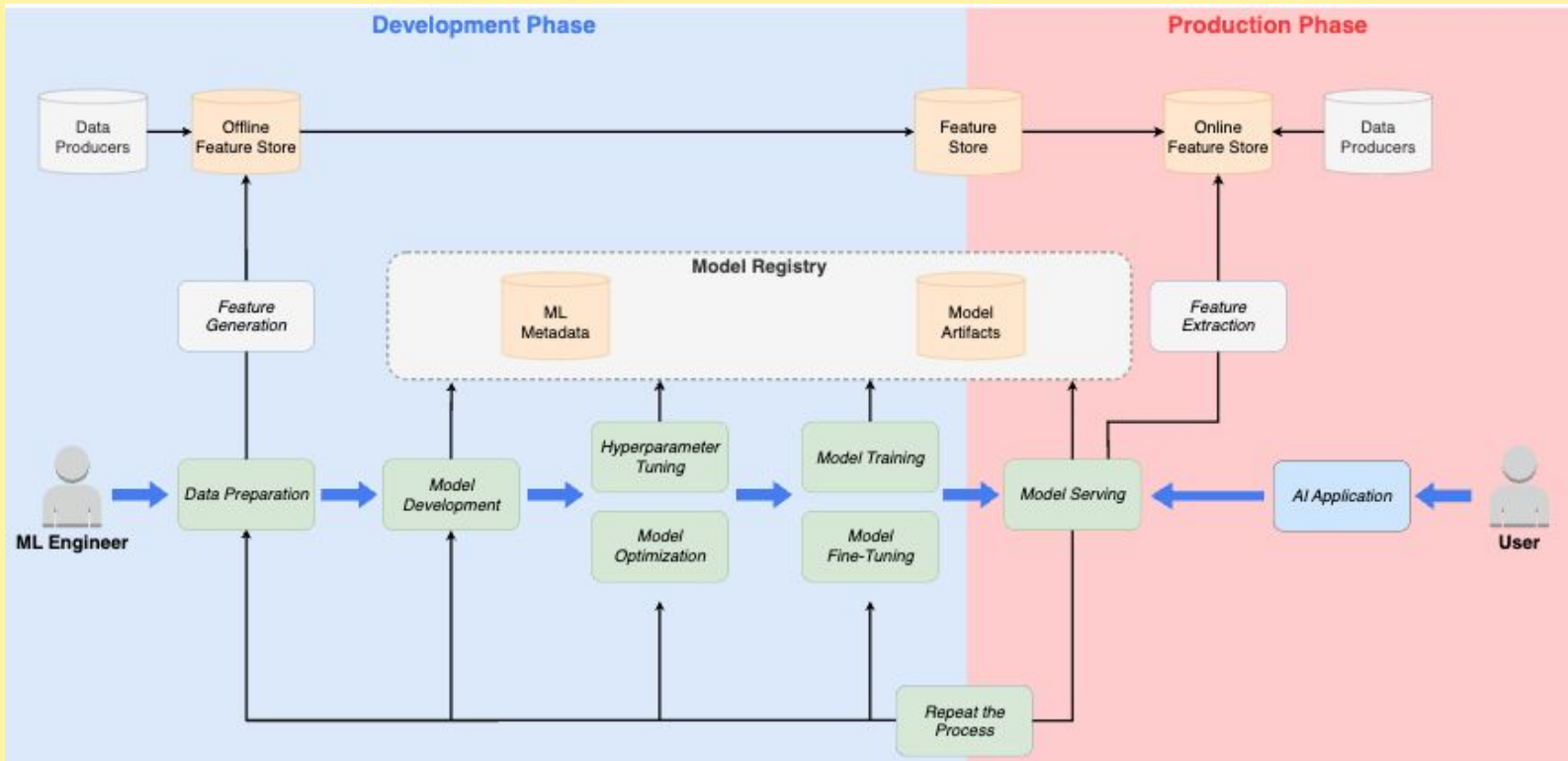
Positive feedback from CERN scientific users

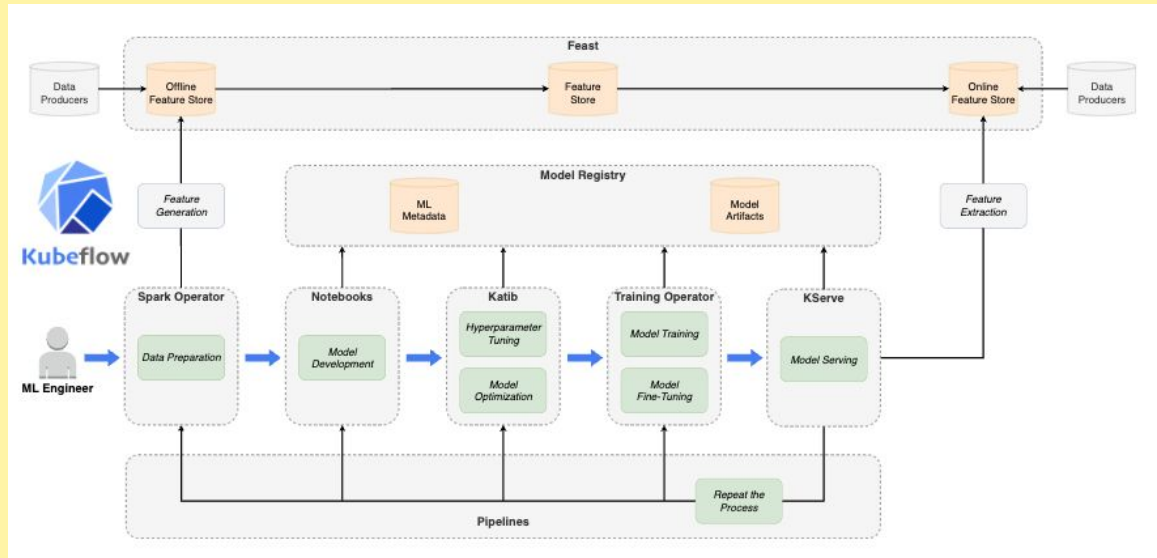
Various use cases running in the cluster

Number of users scaling according to available GPUs

Different Kubeflow features utilized







- [Kubeflow Spark Operator](#) data preparation and feature engineering
- [Kubeflow Notebooks](#) model development and interactive data science to experiment with ML workflows.
- [Kubeflow Katib](#) model optimization and hyperparameter tuning
- [Kubeflow Training Operator](#) large-scale distributed training or fine-tuning
- [Kubeflow Model Registry](#) store ML metadata, model artifacts, and preparing models for production serving.
- [KServe](#) online and batch inference in the model serving step.
- [Feast](#) feature store and to manage offline and online features.
- [Kubeflow Pipelines](#) build, deploy, and manage each step in the ML lifecycle.

05




Local AI



slides



Local AI

- Local, OpenAI drop-in alternative REST API. You own your data.  158
- NO GPU required. NO Internet access is required either  1.7k
- Optional, GPU Acceleration is available. See also the build section.
- Supports multiple models  21.8k
- 🏃 Once loaded the first time, it keep models loaded in memory for faster inference
- ⚡ Doesn't shell-out, but uses bindings for a faster inference and better performance.

Run with Kubernetes: <https://localai.io/basics/kubernetes/>

```
kubectl apply -f https://raw.githubusercontent.com/mudler/LocalAI/master/examples/kubernetes/deployment.yaml
```

<https://localai.io/>

The screenshot shows a web browser window with the URL `localhost:8080/chat/gpt-4`. The page title is "LocalAI" and the navigation bar includes "Home", "Documentation", "Models", "Chat", "Generate images", "TTS", and "API". The chat interface has three buttons: "SET SYSTEM PROMPT", "SET API KEY", and "CLEAR CHAT", along with a "Select a model" dropdown.

User:
what is kcd istanbul about?

Assistant (gpt-4):
KCD Istanbul is an international design festival held in Istanbul, Turkey, dedicated to promoting design, creativity, and innovation. The festival aims to bring together local and international designers, brands, manufacturers, and design enthusiasts to showcase their work, exchange ideas, and celebrate design in various fields such as fashion, graphics, industrial design, and architecture. Through a series of events, exhibitions, workshops, and conferences, KCD Istanbul provides a platform for participants to network, learn from each other, and inspire new design ideas. The festival helps foster a design culture in Istanbul and contributes to the growth of the local creative industry.

User:
hmmm that's not correct. I mean the Kubernetes Community Event. Do you know when it happens?

Assistant (gpt-4):
Apologies for the confusion. I misunderstood your initial question. The Kubernetes Community Days (KCD) events are indeed related to Kubernetes, a popular open-source container orchestration platform.
Kubernetes Community Events are gatherings of Kubernetes enthusiasts, developers, users, and contributors who come together to share knowledge, learn from each other, and collaborate on various aspects of the platform. These events can take different forms, such as conferences, hackathons, meetups, and workshops, and are organized by the local Kubernetes community and official Kubernetes partners.
KCD Istanbul, in particular, is an instance of a Kubernetes Community Event held in Istanbul, Turkey. You can find the dates and further information about specific KCD events on the official Kubernetes website or social media channels. These events usually take place periodically throughout the year and provide an excellent opportunity for participants to engage with the Kubernetes community, share experiences, and learn about the latest updates and best practices in container orchestration and cloud-native applications.

Send a message...

Credits Tobias Schneck

01 **Kubernetes** -> Infrastructure orchestration and management

02 **Kubectl-ai** -> Kubernetes command-line helper

03 **K8sgpt** -> Kubernetes troubleshooter and analyzer

04 **Kubeflow** -> MLOps platform for Kubernetes

05 **Local AI** -> Local model deployment and serving

slides



05

More info

slides



Kubernetes AI Days

The screenshot shows the top navigation bar with the Kubernetes AI Day Europe logo and menu items: Attend, Program, Contact Us, and View All Events. Below the navigation is a yellow banner with a calendar icon and the text: "This event has passed. View the upcoming KubeCon + CloudNativeCon + other CNCF Events." The main content area features the Kubernetes AI Day Europe logo, the date "16 MAY", the location "VALENCIA, SPAIN", and the hashtag "#k8sAI + #k8sAlday".

The screenshot shows the top navigation bar with the Kubernetes AI Day North America logo and menu items: Attend, Program, Contact Us, and View All Events. Below the navigation is a yellow banner with a calendar icon and the text: "This event has passed. View the upcoming KubeCon + CloudNativeCon + other CNCF Events." The main content area features the Kubernetes AI Day North America logo, the date "OCTOBER 25", the location "DETROIT, MICHIGAN", and the hashtag "#K8SAIDAY".

**Cloud Native +
Kubernetes AI Day**

- About
- Call For Proposals
- Registration Details
- Sponsor the Event

ABOUT

**CLOUD NATIVE &
KUBERNETES
AI DAY
NORTH AMERICA**

NOVEMBER 12, 2024
SALT LAKE CITY, UTAH
#CNK8sAIDay

Cloud Native Events



About Projects Training

WHO WE ARE

The Cloud Native Computing Foundation (CNCF) hosts critical components of the global technology infrastructure.

We bring together the world's top developers, end users, and vendors and run the largest open source developer conferences. CNCF is part of the nonprofit [Linux Foundation](#).

THE LINUX FOUNDATION

Sign In



Events

About | Sponsor | Submit a Talk | Calendar | Code of Conduct

Filter events by name, topic...

KubeCon + CloudNativeCon + oth...

All Countries

Reset

KubeCon + CloudNativeCon + Open Source Summit + AI_dev China

Aug 21-23, 2024 Hong Kong

The Cloud Native Computing Foundation's flagship conference joins with Open Source Summit and AI_dev to gather adopters and technologists from leading open source and cloud native communities in Hong Kong, China from 21-23 August, 2024. Join us and immerse yourself in the heart of cloud native computing, open source innovation, and the intricate realm of generative AI and machine learning.

REGISTER SPONSOR SCHEDULE

KubeDay Japan

Aug 27, 2024 Tokyo, Japan

KubeDay events connect international and local experts in global cities with adopters, developers, and practitioners to promote face-to-face collaboration and deliver rich educational experiences.

REGISTER SPONSOR SCHEDULE

gRPCConf

Aug 27, 2024 Sunnyvale, United States

Join key customers, project leads, and contributors that make up the gRPC ecosystem for a full day of talks, demos, case studies, and code labs.

REGISTER SPONSOR SCHEDULE

PromCon Europe

Sep 11-12, 2024 Berlin, Germany

PromCon aims to connect Prometheus users and developers from around the world in order to exchange knowledge, best practices, and experience gained around using Prometheus. We also want to collaborate to build a community and grow professional connections around systems and service monitoring.

REGISTER SPEAK SPONSOR

<https://events.linuxfoundation.org/>

Do you have any questions?



Ceyda Düzgeç
Cloud & Platform Engineer



Thanks!

slides

