KUBERNETES COMMUNITY DAYS BEIJING 2025

Fine-tuning LLM with Argo Workflows：A Kubernetes-native Approach

Shuangkun Tian

Argo Maintainer、Alibaba Cloud Software Engineer

KCD
GROWING CLOUD NATIVE TOGETHER
BEIJING

Content

KCD
GROWING CLOUD NATIVE TOGETHER
BEIJING

Part 01

The Challenge of Fine-tuning

What is Fine-tuning ?

- Adapts pre-trained models to specific tasks/domains via targeted training.

Base Model: Deepseek R1、GPT-3、BERT

Fine-tuned Model: DeepSeek-Finance、Claude、SciBERT

Challenge of Fine-tuning

- Substantial Computational Resources
  - 💻 Various heterogeneous devices：CPU、GPU、DPU
  - 🔥 Computational Cost：Single training run over $10k+
- Complex Workflows
  - ⛓️ Muti-stages：Preprocessing → Training → Evaluation …
  - ⚙️ Larger Tasks（1w+）、Muti-workflows（1000+）

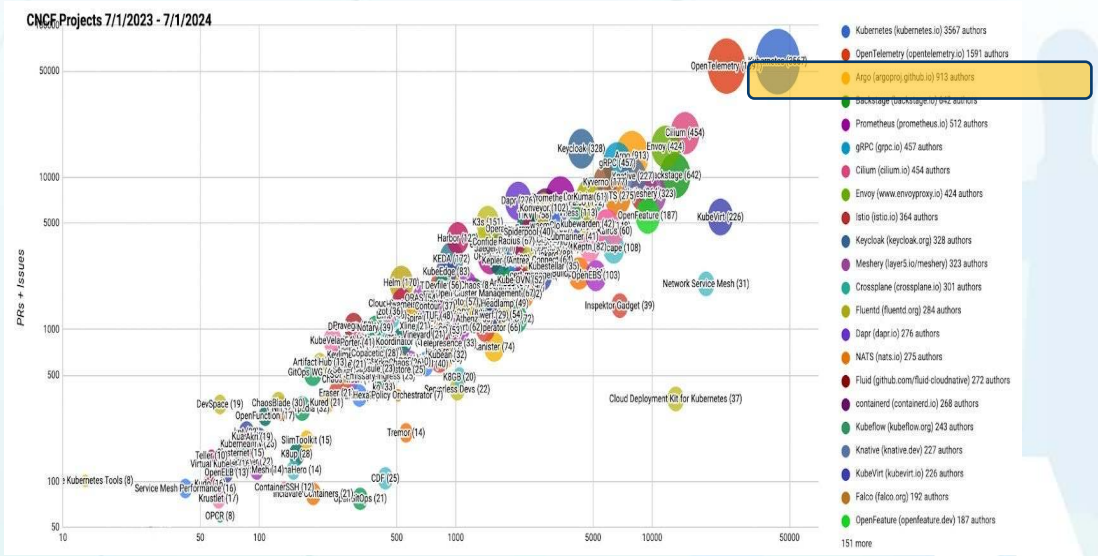
Manual workflows = High cost + Low reliability

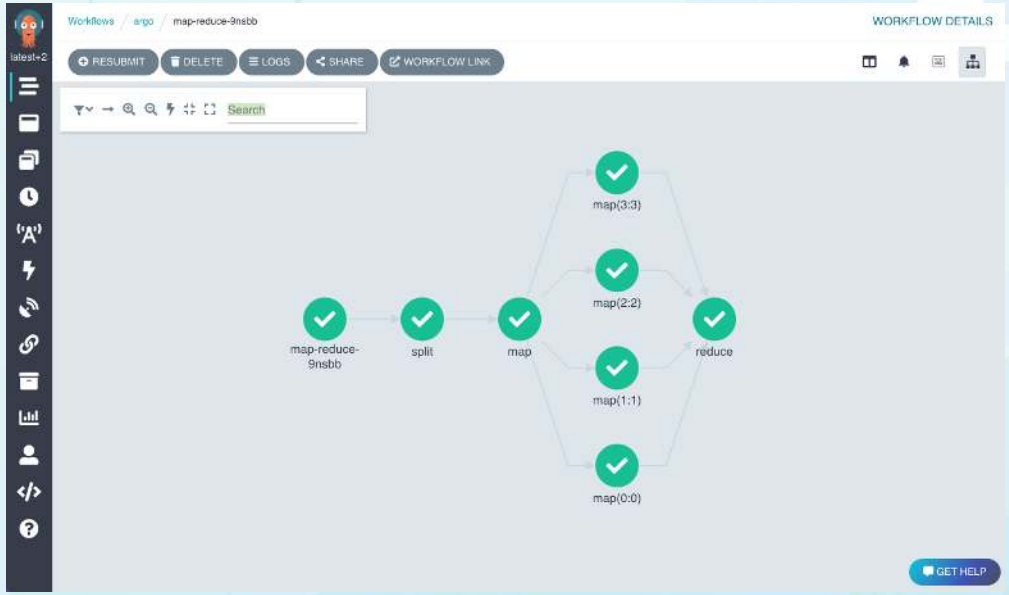
KCD
GROWING CLOUD NATIVE TOGETHER
BEIJING

Part 02
Why Argo Workflows ?

What is Argo Workflows ?

- Argo Project
  - Argo Workflows
  - Argo CD
  - Argo Events、
  - Argo Rollout
  - Third Active Community in CNCF





- Most Popular Workflow Engine
  - Machine Learning pipelines
  - Data and batch processing
  - Infrastructure automation
  - CI/CD

Argo Workflows for ML Pipelines

Kubeflow Pipelines

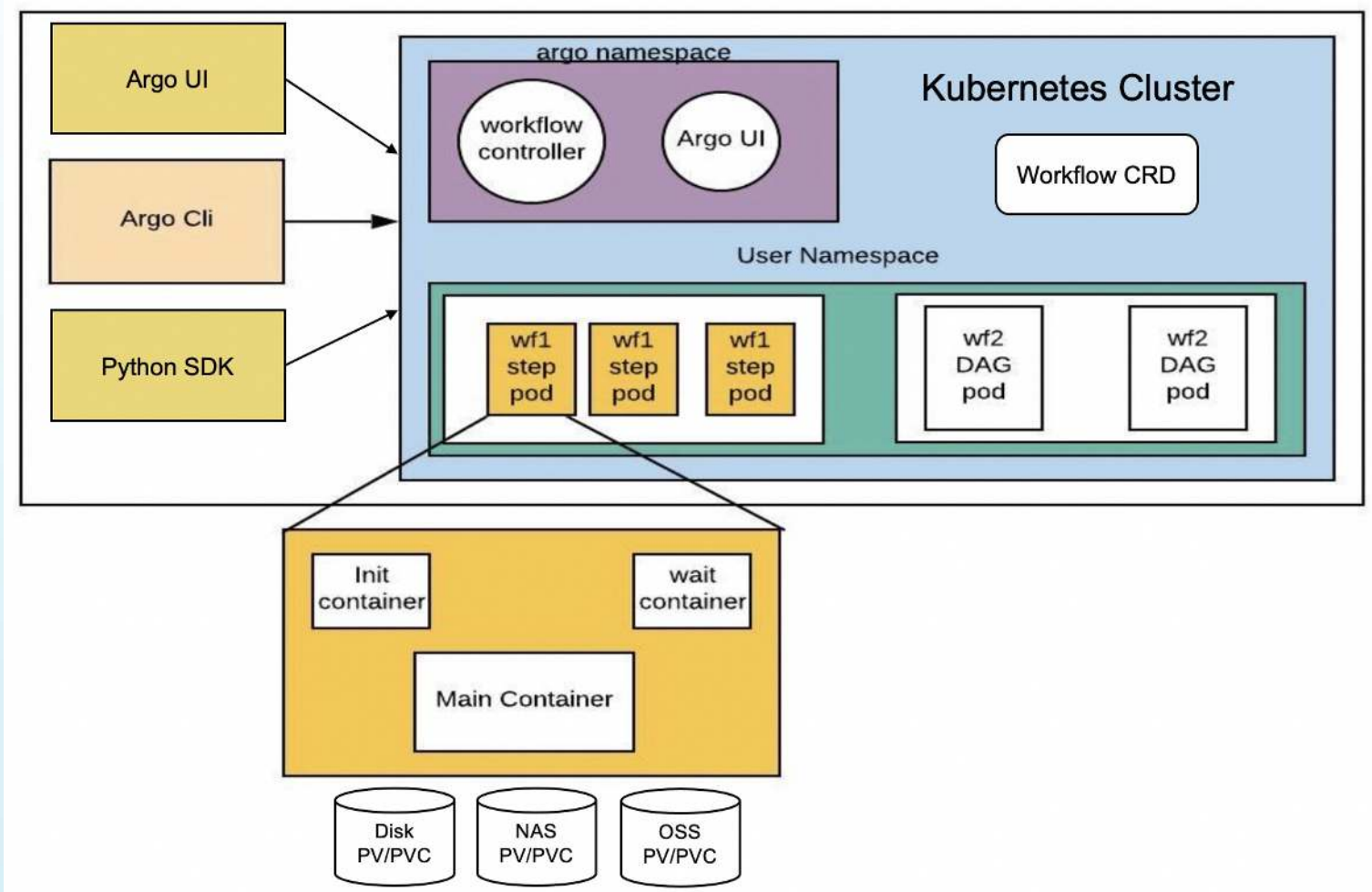
Numa flow

Argo Workflows

- Over 8K companies use Argo or those ML tools based on Argo
- Argo Workflows has been a core component in the orchestration of AI/ML workloads on Kubernetes.

Why Argo Workflows for AI/Fine-tuning

- Kubernetes-native
- Scalability
- Reproducibility
- Fault tolerance
- Visibility
- Ease of Use
- YAML/Python
- ...



KCD
GROWING CLOUD NATIVE TOGETHER
BEIJING

Part 03

Building a Traditional Chinese Medicine Assistant Based on DeepSeek

Workflow Defintion

```yaml
apiVersion: argoproj.io/v1alpha1
kind: Workflow
metadata:
  name: fine-tuning
spec:
  entrypoint: start
  templates:
  - name: start
    steps:
    - - name: training
        template: fine-tune-template
    - - name: evaluation
        template: evaluation-template
  - name: fine-tune-template
    container:
      image: argo-demo/fine-tuning:beta
      command: [python]
      args: ["fine-tune.py","{{model.name}}"]
      resources:
        requests:
          cpu: 2
          memory: 4Gi
           nvidia.com/gpu: 1
  - name: evaluation-template
    container: ...
```

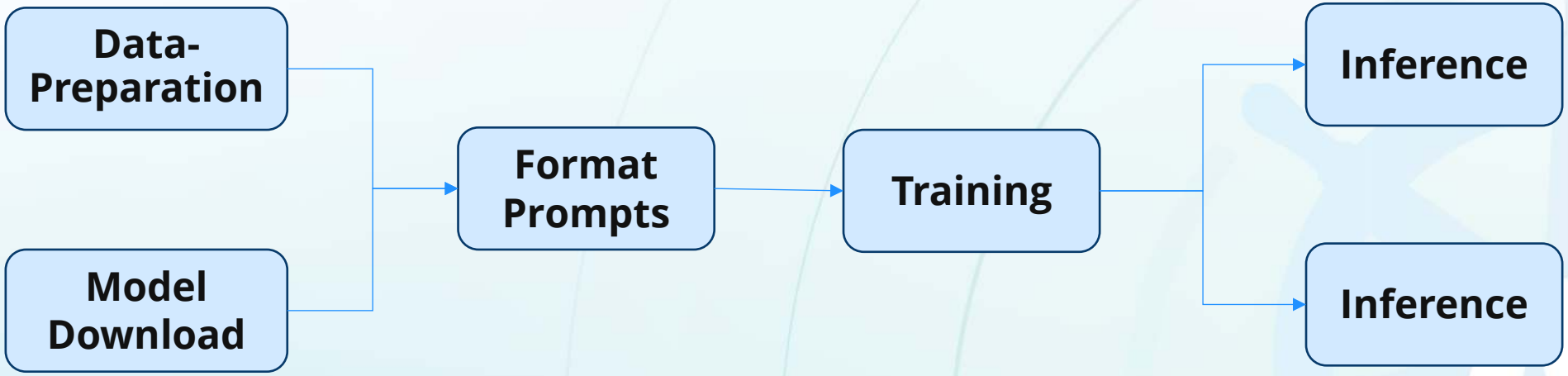
Depenency Definition:
Serial dependencites / Complex DAGs

Template:
Image, Arguments,
Command, and Resouces

Fine-tuning LLM

- Dataset Prepare
    - HuggingFace Hub (e.g., datasets.load_dataset("glue", "marc"))
    - Raw Data → Clean → Tokenize
- Base Model
    - DeepSeek-R1、DeepSeek-R1-Distill
    - 4-bit Quantized Base Model
- Training
    - LoRA Adapters
    - Full Fine-tuning
- Evaluation
    - Human evaluation

Fine-tuning Chinese Medicine Assistant on Deepseek



- Traditional-Chinese-Medicine-Dataset-SFT
- DeepSeek-R1-Distill-Qwen-7B
- 4bit

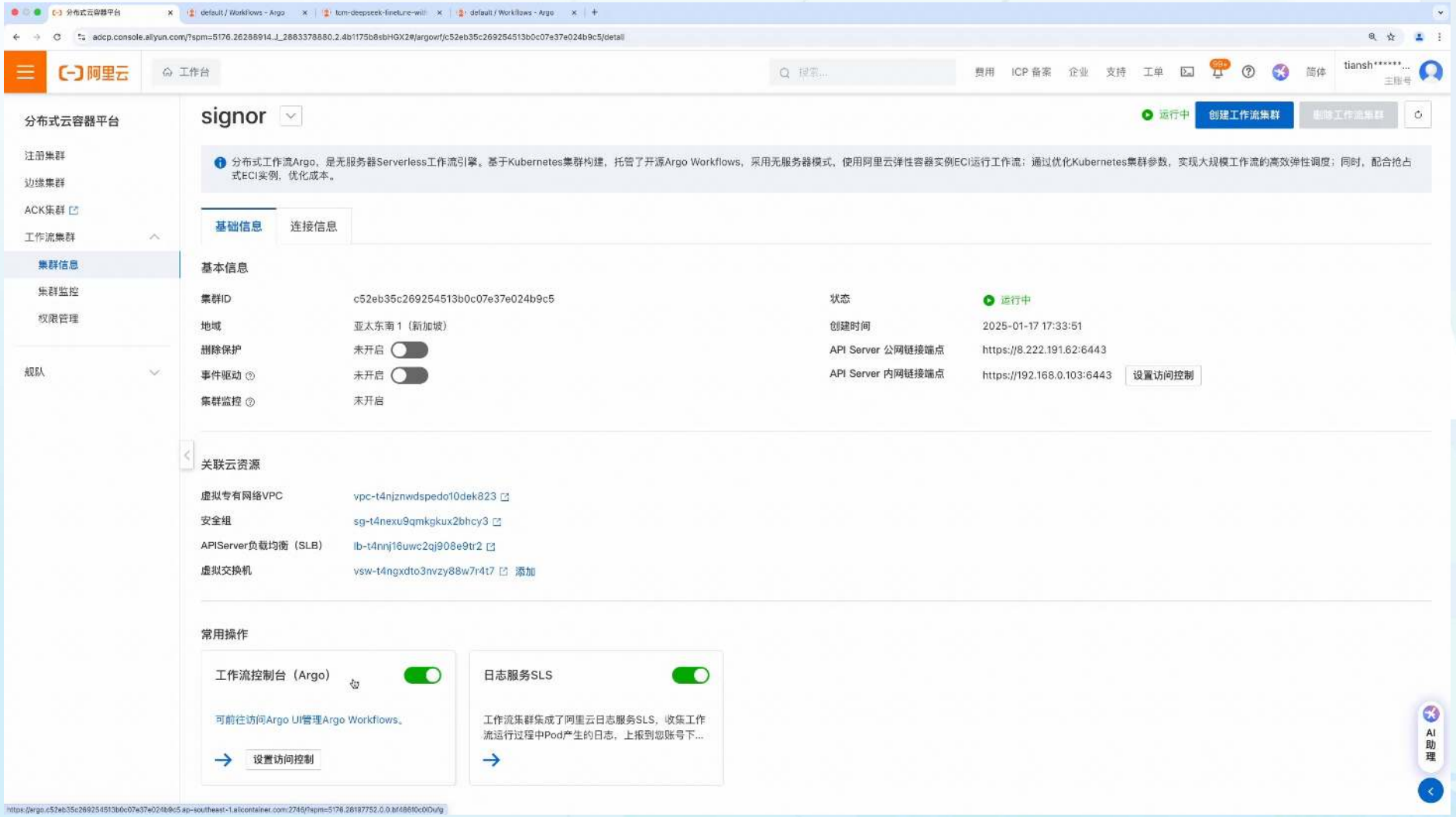
- Tokenization
- Prompt_Style
- Traditional Chinese Medicine Expert

- LoRA
- GPU
- Train Parameters

- Base Model
- Fine tuned Model
- How to cure persistent cough

Demo on ACK Argo Workflows in Alibaba Cloud

- https://github.com/AliyunContainerService/argo-workflow-examples/tree/main/fine-tune-with-argo



Part 04

Benefits and Future

Benefits by this Approach

- Save Cost
    - Use CPU to handle time-consuming tasks such as downloads
    - GPU only for train
- Improve efficiency
    - Automation Task Orchestration
    - Ease to Scale
- Reproducibility
    - Traceable
    - Version Control

Future

- CI/CD System
  - Improve R&D efficiency
- Argo Events
  - Event-Driven Workflow
- Integrate Spark、Ray、Pytorch
  - Unify platform

https://github.com/argoproj/argo-workflows

ACK One & Argo Customer Exchange Group

Thanks.