Kserve 開源貢獻

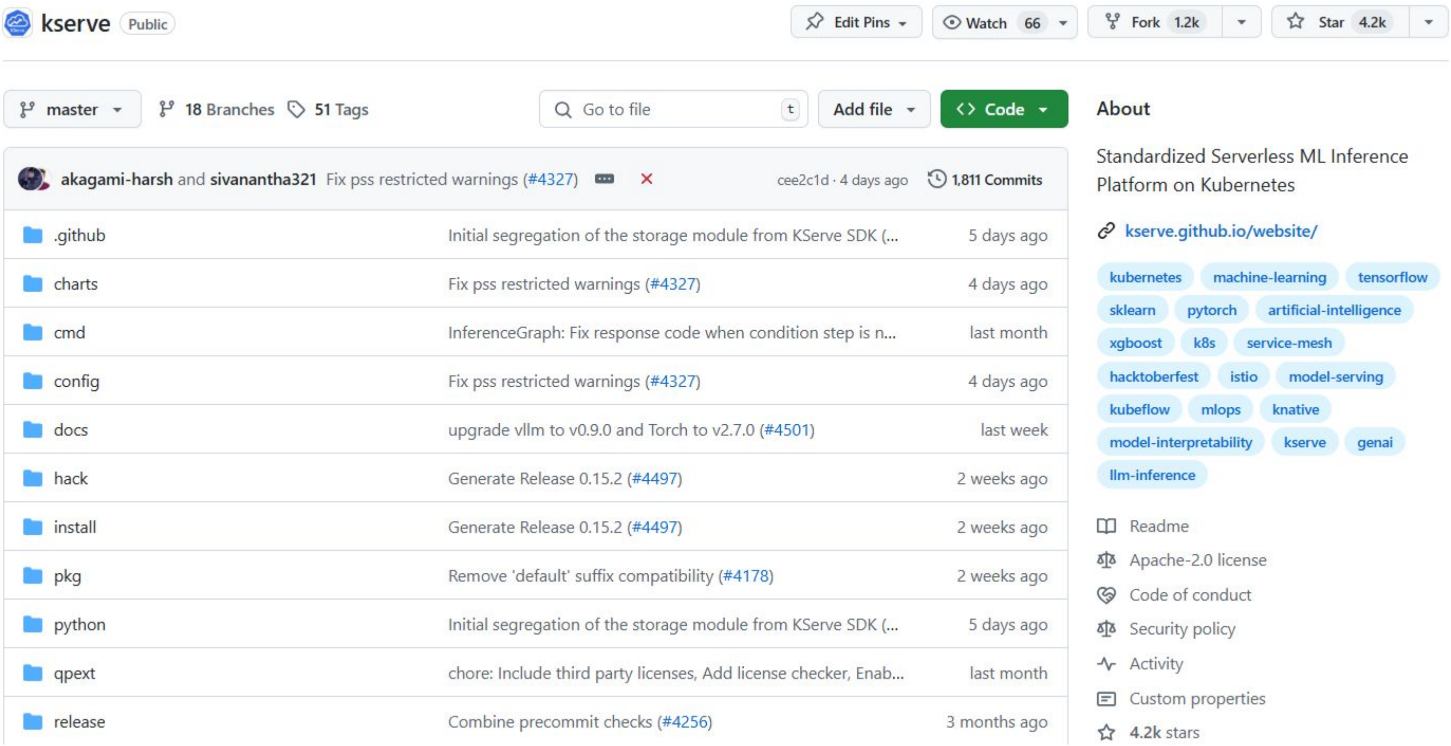
Group 1

kserve是甚麼

🤖 KServe 是什麼？

- KServe 是 在 Kubernetes 上的 AI 模型部署框架
- 幫助我們把訓練好的 AI 模型 ➜ 做成一個 API，能「即時預測」
- 支援多種模型格式：TensorFlow / PyTorch / MLflow / ONNX...

kserve是甚麼



akagami-harsh and sivanantha321 Fix pss restricted warnings (#4327) ✕ cee2c1d · 4 days ago 1,811 Commits

.github	Initial segregation of the storage module from KServe SDK (...	5 days ago
charts	Fix pss restricted warnings (#4327)	4 days ago
cmd	InferenceGraph: Fix response code when condition step is n...	last month
config	Fix pss restricted warnings (#4327)	4 days ago
docs	upgrade vllm to v0.9.0 and Torch to v2.7.0 (#4501)	last week
hack	Generate Release 0.15.2 (#4497)	2 weeks ago
install	Generate Release 0.15.2 (#4497)	2 weeks ago
pkg	Remove 'default' suffix compatibility (#4178)	2 weeks ago
python	Initial segregation of the storage module from KServe SDK (...	5 days ago
qpext	chore: Include third party licenses, Add license checker, Enab...	last month
release	Combine precommit checks (#4256)	3 months ago

About

Standardized Serverless ML Inference Platform on Kubernetes

kserve.github.io/website/

kubernetes machine-learning tensorflow sklearn pytorch artificial-intelligence xgboost k8s service-mesh hacktoberfest istio model-serving kubeflow mlops knative model-interpretability kserve genai llm-inference

- Readme
- Apache-2.0 license
- Code of conduct
- Security policy
- Activity
- Custom properties
- 4.2k stars

開源貢獻是甚麼

💡 什麼是開源（Open Source）？

- 原始碼公開，任何人都能閱讀、使用、修改、貢獻
- 常見平台:GitHub (kserve、openCV)

🔧 可以怎麼貢獻？

- 改進文件
- 回報錯誤（開 issue）
- 寫程式碼（修 bug、新功能）
- 參與討論（Pull Request 審核、留言）

開源貢獻流程

1.尋找適合的問題(ISSUE)

2.在本地端重現問題

3.用各種方式解決問題並跟rewiewers討論

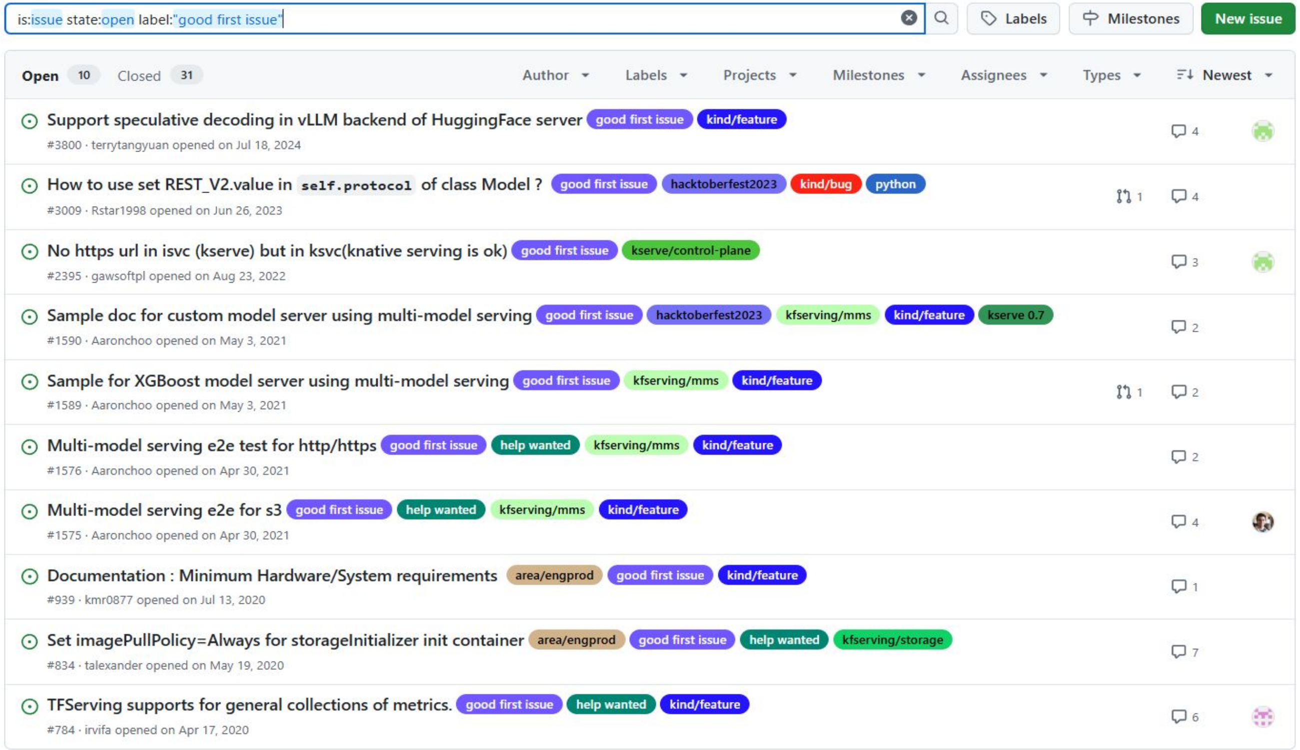
4.提出Pull Request (你自己用甚麼方式解決這個問題)

5.跟rewiewers討論pull request是否能成功合併或需再修改

6.成功被merge(貢獻)

尋找適合問題
並且成功重現

尋找適合問題



is:issue state:open label:"good first issue"

Open 10 Closed 31

Support speculative decoding in vLLM backend of HuggingFace server good first issue kind/feature
#3800 · terrytangyuan opened on Jul 18, 2024 · 4

How to use set REST_V2.value in `self.protocol` of class Model ? good first issue hacktoberfest2023 kind/bug python
#3009 · Rstar1998 opened on Jun 26, 2023 · 1 · 4

No https url in isvc (kserve) but in ksvc(knative serving is ok) good first issue kserve/control-plane
#2395 · gawsoftpl opened on Aug 23, 2022 · 3

Sample doc for custom model server using multi-model serving good first issue hacktoberfest2023 kfserving/mms kind/feature kserve 0.7
#1590 · Aaronchoo opened on May 3, 2021 · 2

Sample for XGBoost model server using multi-model serving good first issue kfserving/mms kind/feature
#1589 · Aaronchoo opened on May 3, 2021 · 1 · 2

Multi-model serving e2e test for http/https good first issue help wanted kfserving/mms kind/feature
#1576 · Aaronchoo opened on Apr 30, 2021 · 2

Multi-model serving e2e for s3 good first issue help wanted kfserving/mms kind/feature
#1575 · Aaronchoo opened on Apr 30, 2021 · 4

Documentation : Minimum Hardware/System requirements area/engprod good first issue kind/feature
#939 · kmr0877 opened on Jul 13, 2020 · 1

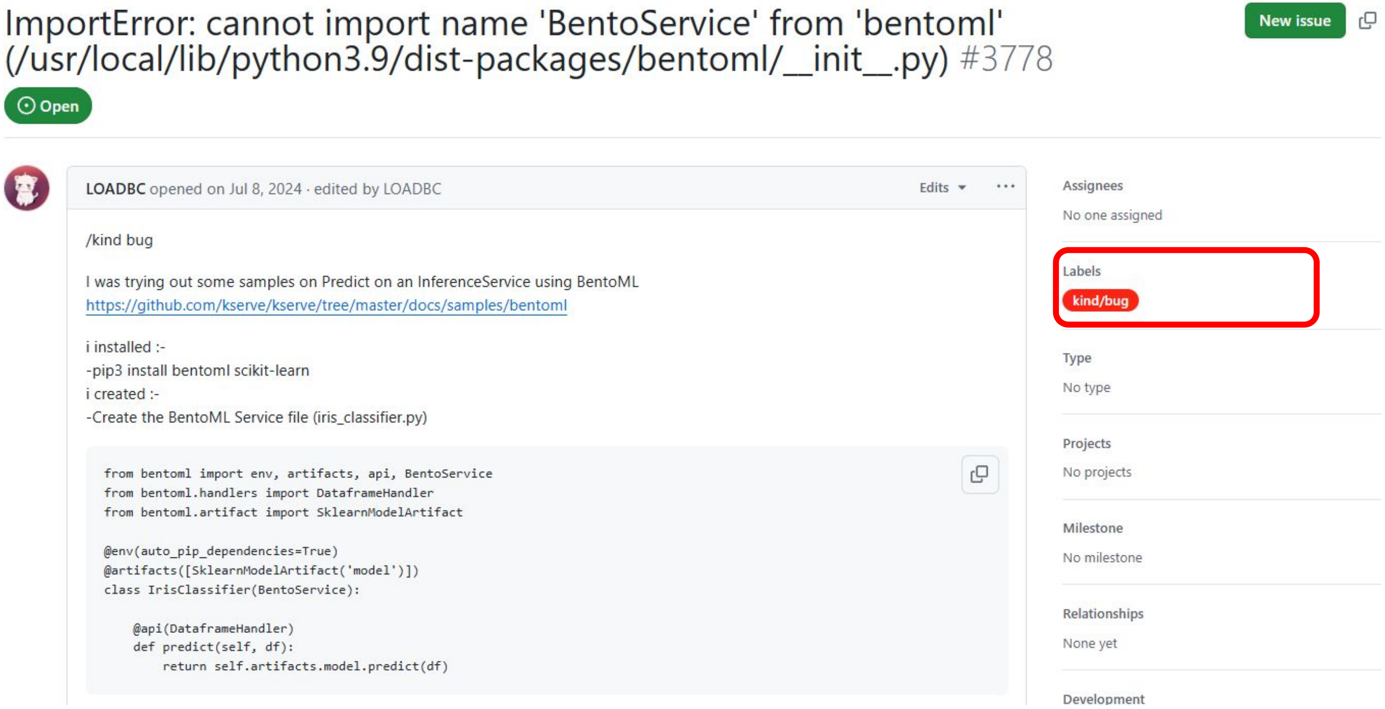
Set imagePullPolicy=Always for storageInitializer init container area/engprod good first issue help wanted kfserving/storage
#834 · talexander opened on May 19, 2020 · 7

TFServing supports for general collections of metrics. good first issue help wanted kind/feature
#784 · irvifa opened on Apr 17, 2020 · 6

GIT指令

- GIT CLONE
- GIT CHECKOUT <BRANCH>
- GIT STATUS
- GIT ADD
- GIT COMMIT
- GIT PUSH

問題重現



ImportError: cannot import name 'BentoService' from 'bentoml' (/usr/local/lib/python3.9/dist-packages/bentoml/__init__.py) #3778

LOADBC opened on Jul 8, 2024 · edited by LOADBC

/kind bug

I was trying out some samples on Predict on an InferenceService using BentoML
https://github.com/kserve/kserve/tree/master/docs/samples/bentoml

i installed :-
-pip3 install bentoml scikit-learn
i created :-
-Create the BentoML Service file (iris_classifier.py)

```
from bentoml import env, artifacts, api, BentoService
from bentoml.handlers import DataframeHandler
from bentoml.artifact import SklearnModelArtifact

@env(auto_pip_dependencies=True)
@artifacts([SklearnModelArtifact('model')])
class IrisClassifier(BentoService):

    @api(DataframeHandler)
    def predict(self, df):
        return self.artifacts.model.predict(df)
```

Assignees
No one assigned

Labels
kind/bug

Type
No type

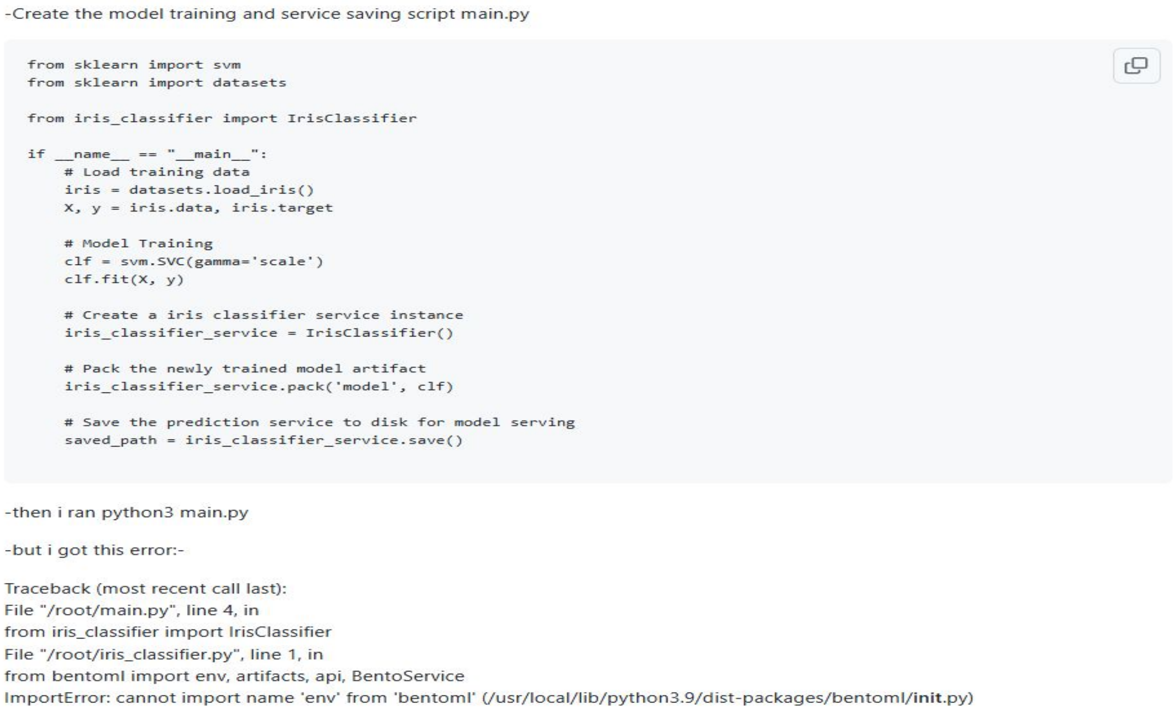
Projects
No projects

Milestone
No milestone

Relationships
None yet

Development

問題重現



-Create the model training and service saving script main.py

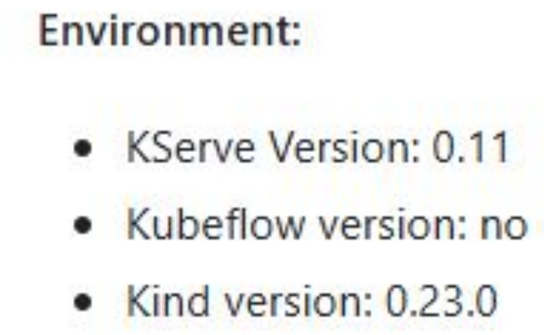
```
from sklearn import svm
from sklearn import datasets

from iris_classifier import IrisClassifier

if __name__ == "__main__":
    # Load training data
    iris = datasets.load_iris()
    X, y = iris.data, iris.target

    # Model Training
    clf = svm.SVC(gamma='scale')
    clf.fit(X, y)

    # Create a iris classifier service instance
    iris_classifier_service = IrisClassifier()

    # Pack the newly trained model artifact
    iris_classifier_service.pack('model', clf)

    # Save the prediction service to disk for model serving
    saved_path = iris_classifier_service.save()
```

-then i ran python3 main.py

-but i got this error:-

```
Traceback (most recent call last):
File "/root/main.py", line 4, in
from iris_classifier import IrisClassifier
File "/root/iris_classifier.py", line 1, in
from bentoml import env, artifacts, api, BentoService
ImportError: cannot import name 'env' from 'bentoml' (/usr/local/lib/python3.9/dist-packages/bentoml/init.py)
```

問題重現

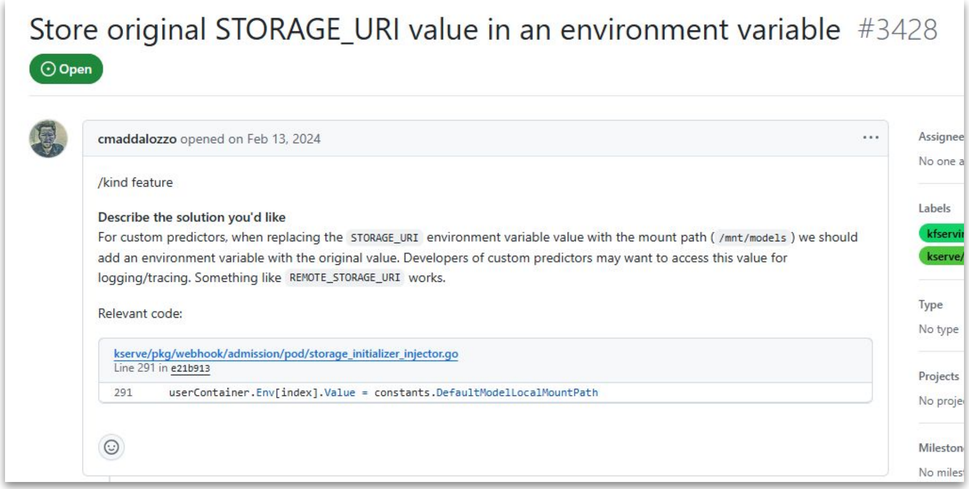


Environment:

- KServe Version: 0.11
- Kubeflow version: no
- Kind version: 0.23.0

解決問題
並且提出PR

Issue #3428 : Store original STORAGE_URI value in an environment variable



Store original STORAGE_URI value in an environment variable #3428

cmaddalozzo opened on Feb 13, 2024

/kind feature

Describe the solution you'd like

For custom predictors, when replacing the `STORAGE_URI` environment variable value with the mount path (`/mnt/models`) we should add an environment variable with the original value. Developers of custom predictors may want to access this value for logging/tracing. Something like `REMOTE_STORAGE_URI` works.

Relevant code:

kserve/pkg/webhook/admission/pod/storage_initializer_injector.go
Line 291 in e21b913

```
291        userContainer.Env[index].Value = constants.DefaultModelLocalMountPath
```

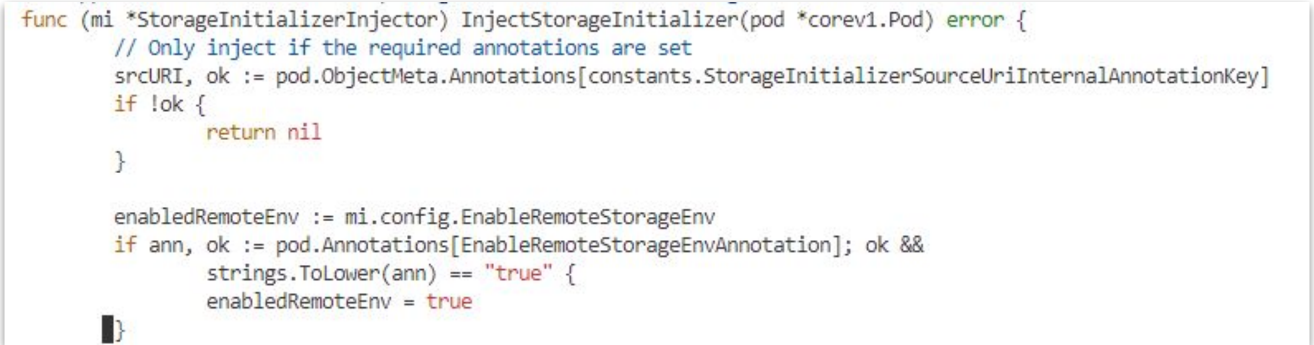
Issue 介紹 :

- STORAGE_URI (環境變數): 告訴 KServe 從哪裡下載模型
- 模型被下載後, KServe 會把這個變數的值換掉, 換成一個本地的路徑
- 希望在 KServe加上一個新的環境變數會保留「原本的路徑值」

Issue #3428 : Store original STORAGE_URI value in an environment variable

storage_initializer_injector.go

- 兩個地方可控制功能
  - 平台預設開關 : 整個平台、所有pod預設行為
  - 個別複寫開關 : 單一pod可自行決定是否開啟

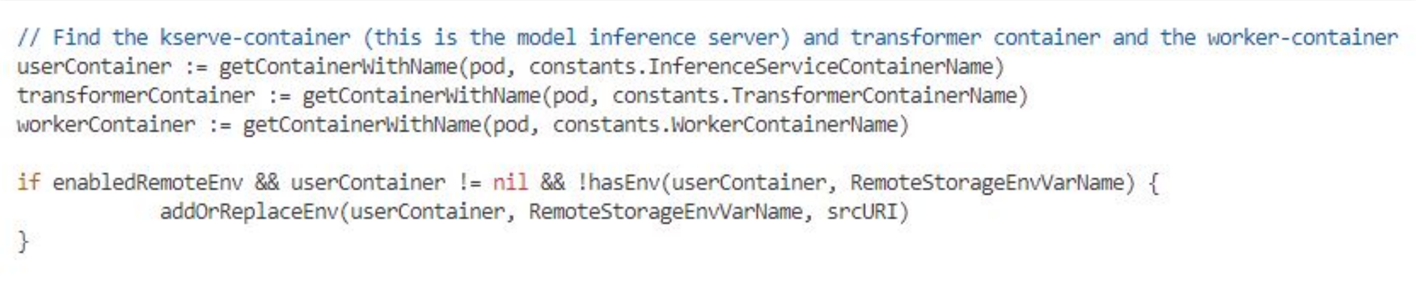


```go
func (mi *StorageInitializerInjector) InjectStorageInitializer(pod *corev1.Pod) error {
	// Only inject if the required annotations are set
	srcURI, ok := pod.ObjectMeta.Annotations[constants.StorageInitializerSourceUriInternalAnnotationKey]
	if !ok {
		return nil
	}

	enabledRemoteEnv := mi.config.EnableRemoteStorageEnv
	if ann, ok := pod.Annotations[EnableRemoteStorageEnvAnnotation]; ok &&
		strings.ToLower(ann) == "true" {
		enabledRemoteEnv = true
	}
```

Issue #3428 : Store original STORAGE_URI value in an environment variable

storage_initializer_injector.go

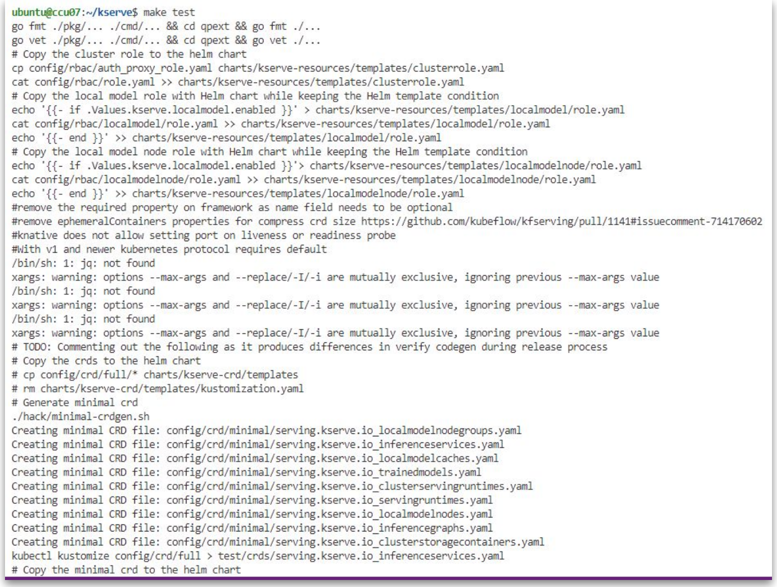
- 注入環境變數 (在找到container後新增)
  - 只在 enabledRemoteEnv == true 時執行。
  - 保證不重覆注入(利用hasEnv 檢查)。



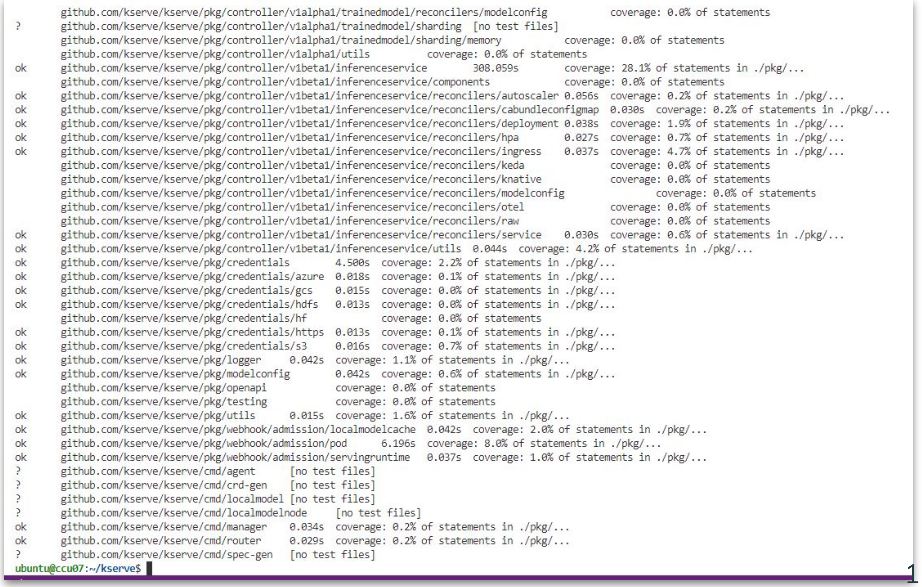
```go
// Find the kserve-container (this is the model inference server) and transformer container and the worker-container
userContainer := getContainerWithName(pod, constants.InferenceServiceContainerName)
transformerContainer := getContainerWithName(pod, constants.TransformerContainerName)
workerContainer := getContainerWithName(pod, constants.WorkerContainerName)

if enabledRemoteEnv && userContainer != nil && !hasEnv(userContainer, RemoteStorageEnvVarName) {
        addOrReplaceEnv(userContainer, RemoteStorageEnvVarName, srcURI)
}
```

Issue #3428 : Store original STORAGE_URI value in an environment variable

- make test 測試已通過



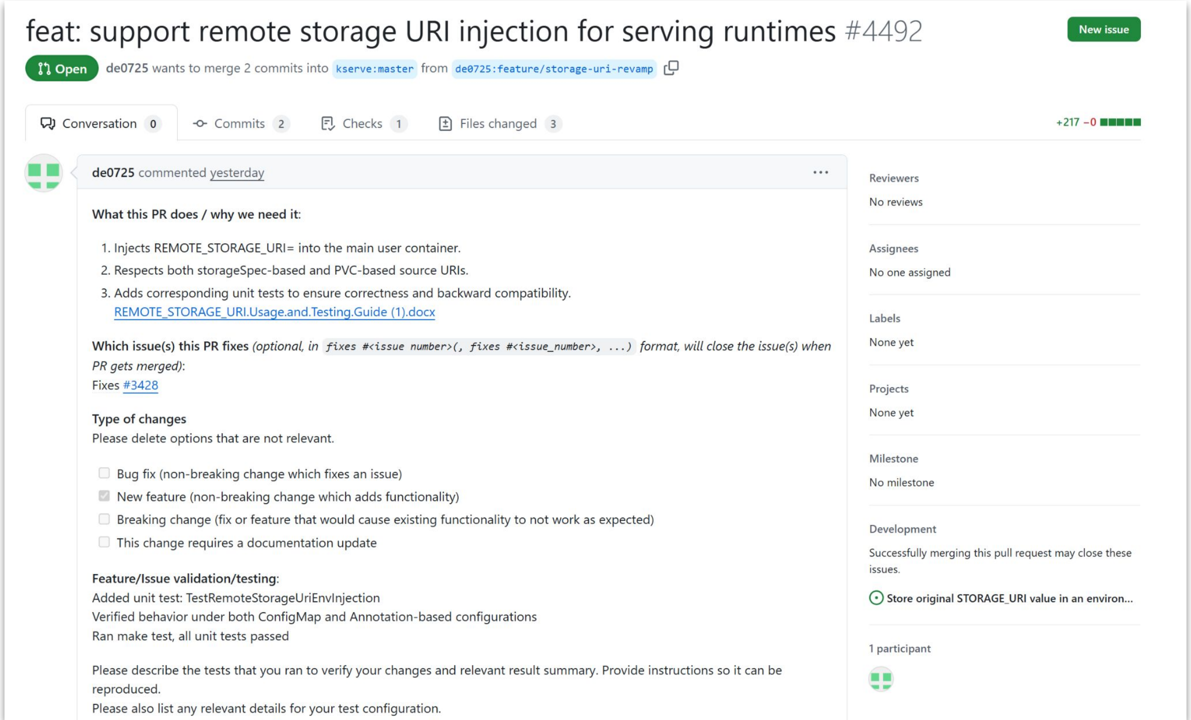
```
ubuntu@ccu07:~/kserve$ make test
go fmt ./pkg/... ./cmd/... && cd qpext && go fmt ./...
go vet ./pkg/... ./cmd/... && cd qpext && go vet ./...
# Copy the cluster role to the helm chart
cp config/rbac/auth_proxy_role.yaml charts/kserve-resources/templates/clusterrole.yaml
cat config/rbac/role.yaml >> charts/kserve-resources/templates/clusterrole.yaml
# Copy the local model role with Helm chart while keeping the Helm template condition
echo '{{- if .Values.kserve.localmodel.enabled }}' > charts/kserve-resources/templates/localmodel/role.yaml
cat config/rbac/localmodel/role.yaml >> charts/kserve-resources/templates/localmodel/role.yaml
echo '{{- end }}' >> charts/kserve-resources/templates/localmodel/role.yaml
# Copy the local model node role with Helm chart while keeping the Helm template condition
echo '{{- if .Values.kserve.localmodel.enabled }}'> charts/kserve-resources/templates/localmodelnode/role.yaml
cat config/rbac/localmodelnode/role.yaml >> charts/kserve-resources/templates/localmodelnode/role.yaml
echo '{{- end }}' >> charts/kserve-resources/templates/localmodelnode/role.yaml
#remove the required property on framework as name field needs to be optional
#remove ephemeralContainers properties for compress crd size https://github.com/kubeflow/kfserving/pull/1141#issuecomment-714170602
#knative does not allow setting port on liveness or readiness probe
#With v1 and newer kubernetes protocol requires default
/bin/sh: 1: jq: not found
xargs: warning: options --max-args and --replace/-I/-i are mutually exclusive, ignoring previous --max-args value
/bin/sh: 1: jq: not found
xargs: warning: options --max-args and --replace/-I/-i are mutually exclusive, ignoring previous --max-args value
/bin/sh: 1: jq: not found
xargs: warning: options --max-args and --replace/-I/-i are mutually exclusive, ignoring previous --max-args value
# TODO: Commenting out the following as it produces differences in verify codegen during release process
# Copy the crds to the helm chart
# cp config/crd/full/* charts/kserve-crd/templates
# rm charts/kserve-crd/templates/kustomization.yaml
# Generate minimal crd
./hack/minimal-crdgen.sh
Creating minimal CRD file: config/crd/minimal/serving.kserve.io_localmodelnodegroups.yaml
Creating minimal CRD file: config/crd/minimal/serving.kserve.io_inferenceservices.yaml
Creating minimal CRD file: config/crd/minimal/serving.kserve.io_localmodelcaches.yaml
Creating minimal CRD file: config/crd/minimal/serving.kserve.io_trainedmodels.yaml
Creating minimal CRD file: config/crd/minimal/serving.kserve.io_clusterservingruntimes.yaml
Creating minimal CRD file: config/crd/minimal/serving.kserve.io_servingruntimes.yaml
Creating minimal CRD file: config/crd/minimal/serving.kserve.io_localmodelnodes.yaml
Creating minimal CRD file: config/crd/minimal/serving.kserve.io_inferencegraphs.yaml
Creating minimal CRD file: config/crd/minimal/serving.kserve.io_clusterstoragecontainers.yaml
kubectl kustomize config/crd/full > test/crds/serving.kserve.io_inferenceservices.yaml
# Copy the minimal crd to the helm chart
```



```
        github.com/kserve/kserve/pkg/controller/v1alpha1/trainedmodel/reconcilers/modelconfig          coverage: 0.0% of statements
?       github.com/kserve/kserve/pkg/controller/v1alpha1/trainedmodel/sharding  [no test files]
        github.com/kserve/kserve/pkg/controller/v1alpha1/trainedmodel/sharding/memory          coverage: 0.0% of statements
        github.com/kserve/kserve/pkg/controller/v1alpha1/utils          coverage: 0.0% of statements
ok      github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice        308.059s        coverage: 28.1% of statements in ./pkg/...
        github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/components             coverage: 0.0% of statements
ok      github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/reconcilers/autoscaler 0.056s  coverage: 0.2% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/reconcilers/cabundleconfigmap  0.030s  coverage: 0.2% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/reconcilers/deployment 0.038s  coverage: 1.9% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/reconcilers/hpa        0.027s  coverage: 0.7% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/reconcilers/ingress    0.037s  coverage: 4.7% of statements in ./pkg/...
        github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/reconcilers/keda               coverage: 0.0% of statements
        github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/reconcilers/knative            coverage: 0.0% of statements
        github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/reconcilers/modelconfig                coverage: 0.0% of statements
        github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/reconcilers/otel               coverage: 0.0% of statements
        github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/reconcilers/raw                coverage: 0.0% of statements
ok      github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/reconcilers/service    0.030s  coverage: 0.6% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/controller/v1beta1/inferenceservice/utils  0.044s  coverage: 4.2% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/credentials        4.500s  coverage: 2.2% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/credentials/azure  0.018s  coverage: 0.1% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/credentials/gcs    0.015s  coverage: 0.0% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/credentials/hdfs   0.013s  coverage: 0.0% of statements in ./pkg/...
        github.com/kserve/kserve/pkg/credentials/hf             coverage: 0.0% of statements
ok      github.com/kserve/kserve/pkg/credentials/https  0.013s  coverage: 0.1% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/credentials/s3     0.016s  coverage: 0.7% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/logger     0.042s  coverage: 1.1% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/modelconfig        0.042s  coverage: 0.6% of statements in ./pkg/...
        github.com/kserve/kserve/pkg/openapi            coverage: 0.0% of statements
        github.com/kserve/kserve/pkg/testing            coverage: 0.0% of statements
ok      github.com/kserve/kserve/pkg/utils      0.015s  coverage: 1.6% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/webhook/admission/localmodelcache  0.042s  coverage: 2.0% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/webhook/admission/pod      6.196s  coverage: 8.0% of statements in ./pkg/...
ok      github.com/kserve/kserve/pkg/webhook/admission/servingruntime   0.037s  coverage: 1.0% of statements in ./pkg/...
?       github.com/kserve/kserve/cmd/agent      [no test files]
?       github.com/kserve/kserve/cmd/crd-gen    [no test files]
?       github.com/kserve/kserve/cmd/localmodel [no test files]
?       github.com/kserve/kserve/cmd/localmodelnode     [no test files]
ok      github.com/kserve/kserve/cmd/manager    0.034s  coverage: 0.2% of statements in ./pkg/...
ok      github.com/kserve/kserve/cmd/router     0.029s  coverage: 0.2% of statements in ./pkg/...
?       github.com/kserve/kserve/cmd/spec-gen   [no test files]
ubuntu@ccu07:~/kserve$
```

Issue #3428 : Store original STORAGE_URI value in an environment variable

- 提出PR



feat: support remote storage URI injection for serving runtimes #4492

de0725 wants to merge 2 commits into `kserve:master` from `de0725:feature/storage-uri-revamp`

de0725 commented yesterday

What this PR does / why we need it:

1. Injects REMOTE_STORAGE_URI= into the main user container.
2. Respects both storageSpec-based and PVC-based source URIs.
3. Adds corresponding unit tests to ensure correctness and backward compatibility.
   REMOTE_STORAGE_URI.Usage.and.Testing.Guide (1).docx

Which issue(s) this PR fixes *(optional, in* `fixes #<issue number>(, fixes #<issue_number>, ...)` *format, will close the issue(s) when PR gets merged)*:
Fixes #3428

Type of changes
Please delete options that are not relevant.

- □ Bug fix (non-breaking change which fixes an issue)
- ☑ New feature (non-breaking change which adds functionality)
- □ Breaking change (fix or feature that would cause existing functionality to not work as expected)
- □ This change requires a documentation update

Feature/Issue validation/testing:
Added unit test: TestRemoteStorageUriEnvInjection
Verified behavior under both ConfigMap and Annotation-based configurations
Ran make test, all unit tests passed

Please describe the tests that you ran to verify your changes and relevant result summary. Provide instructions so it can be reproduced.
Please also list any relevant details for your test configuration.

Reviewers
No reviews

Assignees
No one assigned

Labels
None yet

Projects
None yet

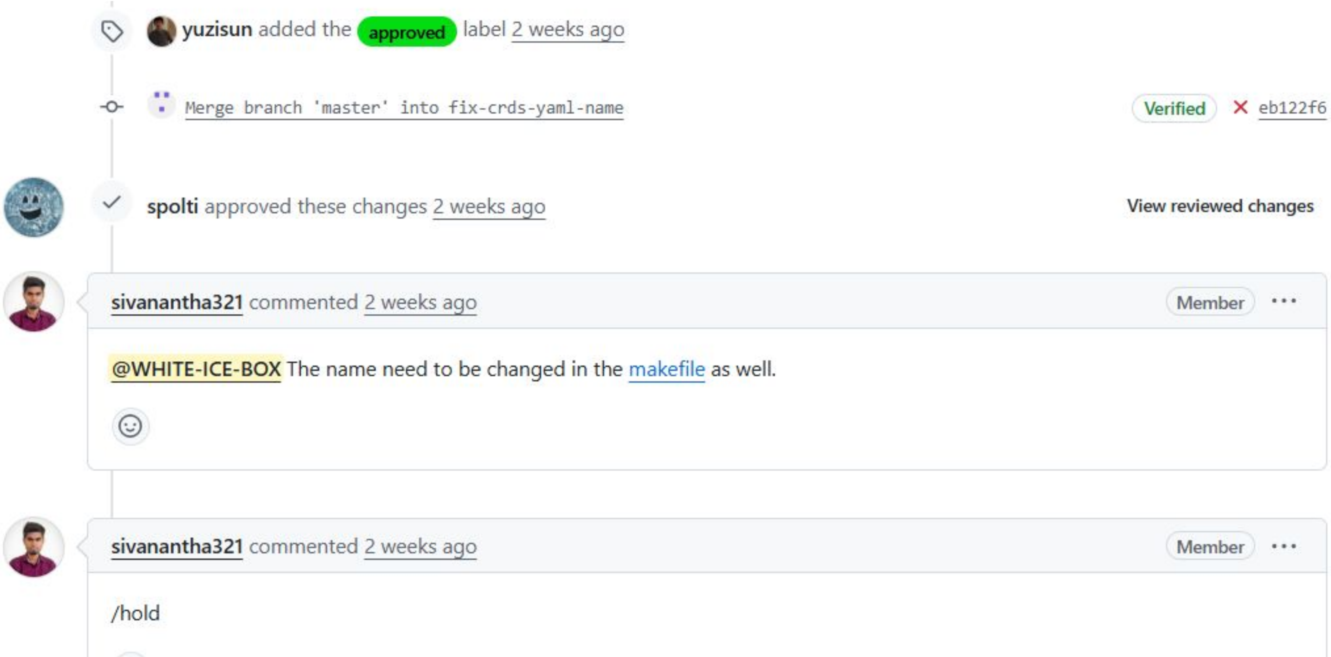
Milestone
No milestone

Development
Successfully merging this pull request may close these issues.
○ Store original STORAGE_URI value in an environ...

1 participant

討論PR可行性
並且成功被MERGE

Rename CRD file to reflect all KServe CRDs (Fixes #4396) #4494



yuzisun added the approved label 2 weeks ago

Merge branch 'master' into fix-crds-yaml-name

Verified ✕ eb122f6

spolti approved these changes 2 weeks ago

View reviewed changes

sivanantha321 commented 2 weeks ago

Member ···

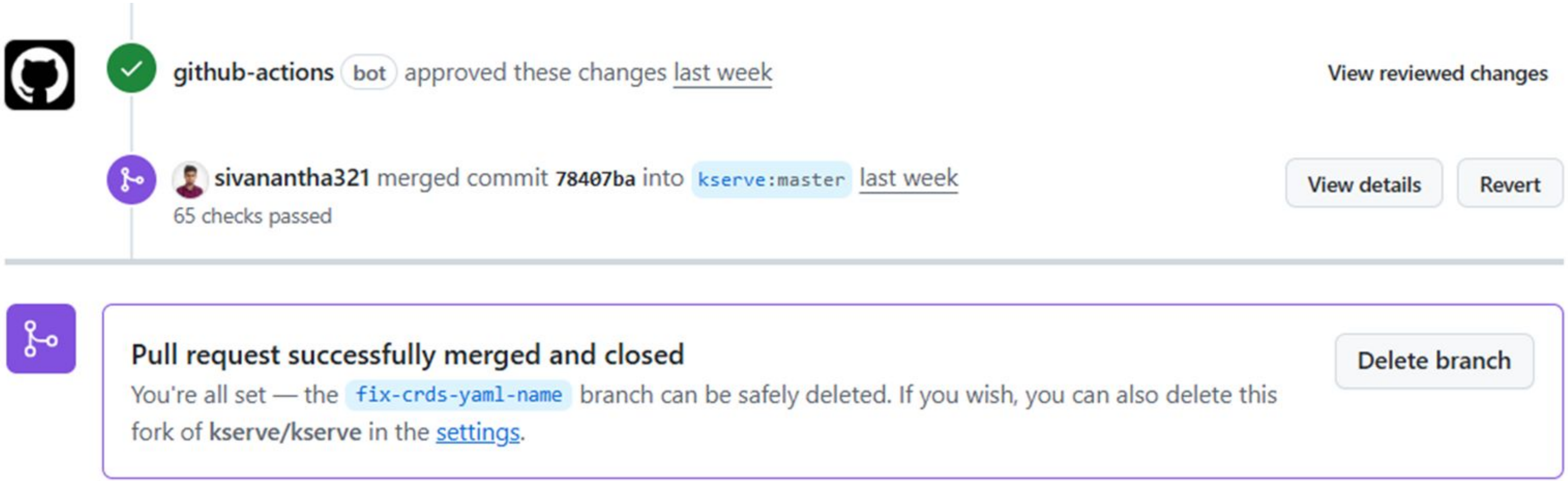
@WHITE-ICE-BOX The name need to be changed in the makefile as well.

sivanantha321 commented 2 weeks ago

Member ···

/hold

Rename CRD file to reflect all KServe CRDs (Fixes #4396) #4494



github-actions bot approved these changes last week

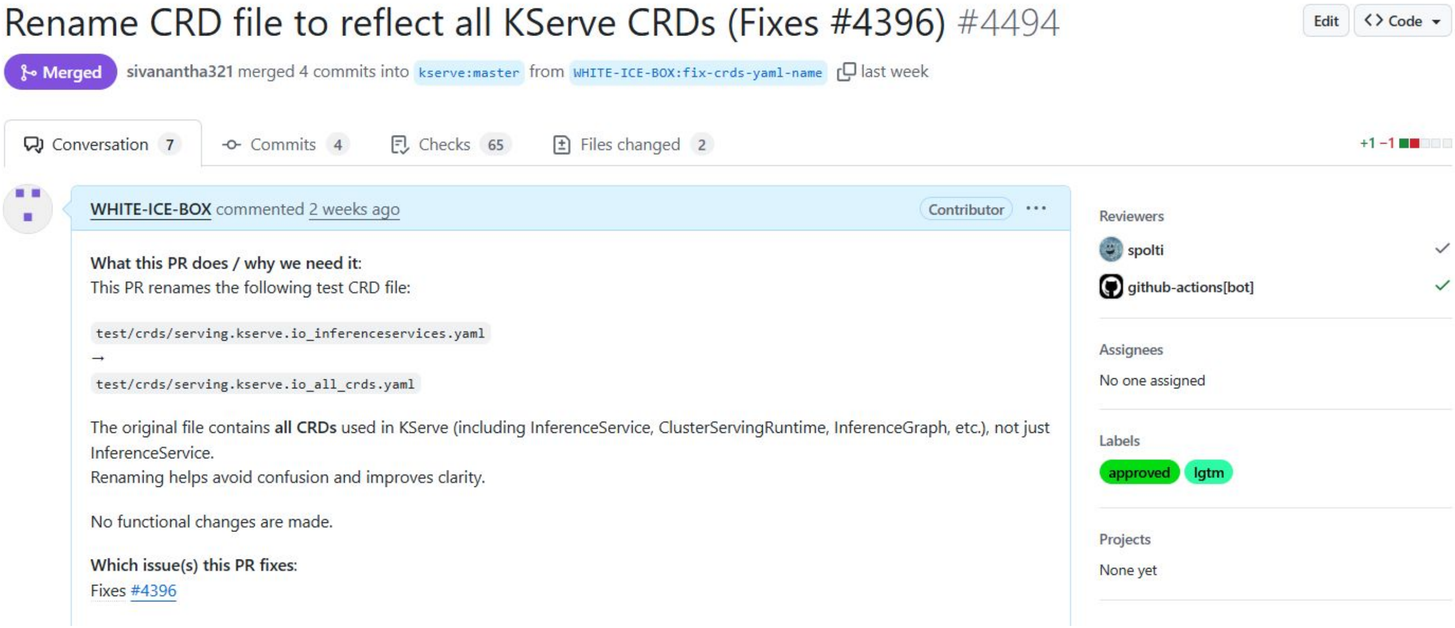
sivanantha321 merged commit 78407ba into kserve:master last week

65 checks passed

Pull request successfully merged and closed

You're all set — the fix-crds-yaml-name branch can be safely deleted. If you wish, you can also delete this fork of **kserve/kserve** in the settings.

Rename CRD file to reflect all KServe CRDs (Fixes #4396) #4494



Rename CRD file to reflect all KServe CRDs (Fixes #4396) #4494

Merged sivanantha321 merged 4 commits into `kserve:master` from `WHITE-ICE-BOX:fix-crds-yaml-name` last week

WHITE-ICE-BOX commented 2 weeks ago

What this PR does / why we need it:
This PR renames the following test CRD file:

`test/crds/serving.kserve.io_inferenceservices.yaml`
→
`test/crds/serving.kserve.io_all_crds.yaml`

The original file contains **all CRDs** used in KServe (including InferenceService, ClusterServingRuntime, InferenceGraph, etc.), not just InferenceService.
Renaming helps avoid confusion and improves clarity.

No functional changes are made.

Which issue(s) this PR fixes:
Fixes #4396

Reviewers
spolti ✓
github-actions[bot] ✓

Assignees
No one assigned

Labels
approved lgtm

Projects
None yet

QNA

🧠 **為什麼要貢獻？**

- 學技術, 練 Git
- 建立個人作品集
- 跟世界各地的高手合作