AI Explainability WG

June 2023

June 2023 updates

- TrustyAI service
  - Support for KServe/ModelMesh "raw" contents[1]
  - Improved ODH e2e testing
  - Exposed explainability endpoints[2]
- Core
  - Refactored TrustyAI metrics namespaces[3]
  - Implemented TrustyAI dataframe metadata (*cf. time-series support*)
- Python TrustyAI
  - Refactored metrics modules
  - Dependencies updates
- Operator
  - TrustyAI service Kubernetes operator[4]

[1]: https://github.com/trustyai-explainability/trustyai-explainability/pull/171
[2]: https://github.com/trustyai-explainability/trustyai-explainability/pull/166
[3]: https://github.com/trustyai-explainability/trustyai-explainability/pull/164
[4]: https://github.com/trustyai-explainability/trustyai-service-operator

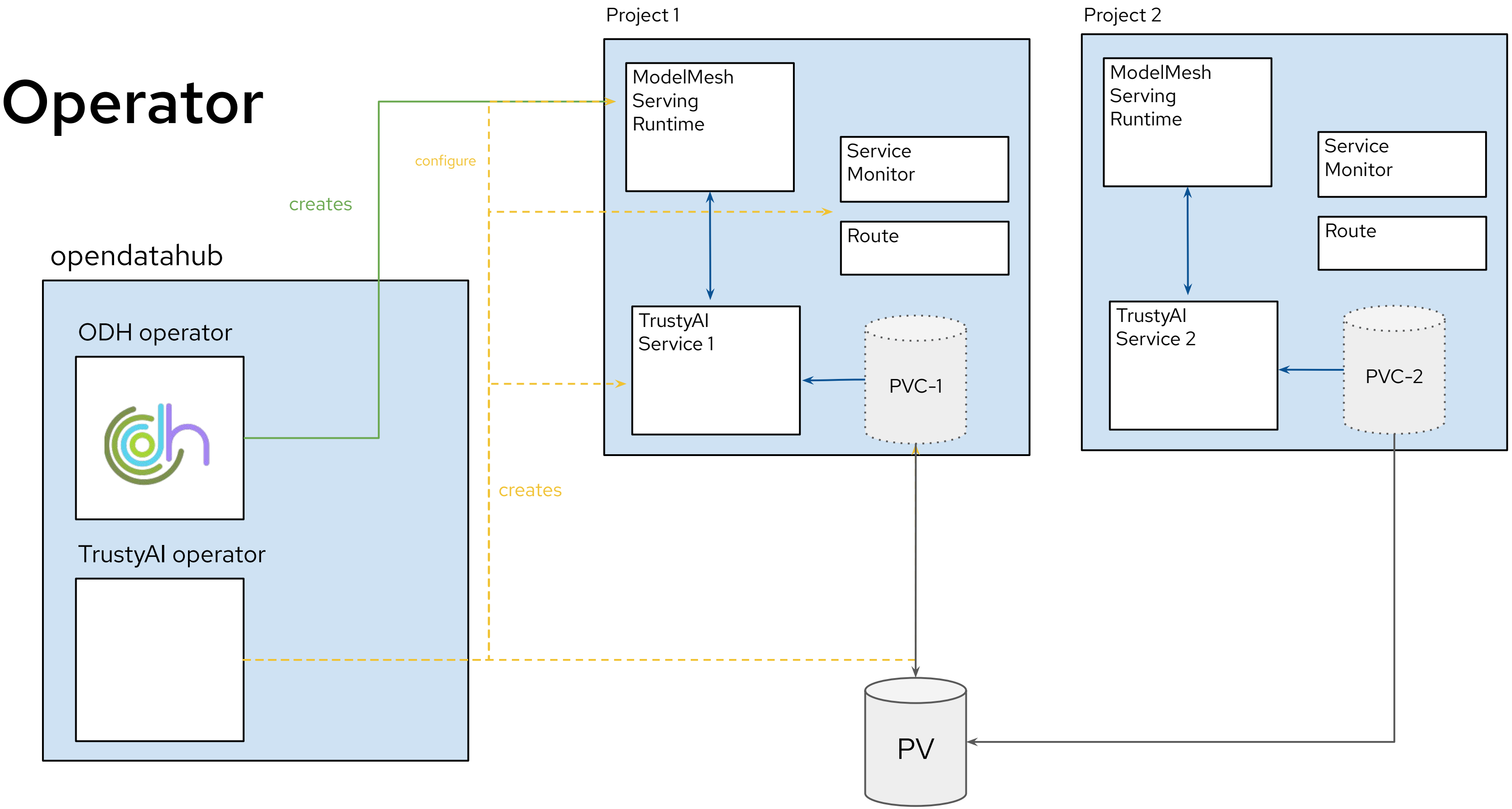
Operator

Operator

- Repositories
    - Operator: https://github.com/trustyai-explainability/trustyai-service-operator
    - ADR-0003: https://github.com/trustyai-explainability/community/pull/5
- Responsibilities
    - Deploying `TrustyAIService` instances
    - Managing service monitors
    - Managing Routes
    - Creating and managing storage[1]
    - Registering TrustyAI as a ModelMesh payload processor
    - Simple CR

[1]: Only PVC storage supported at the moment

Operator



opendatahub

ODH operator

TrustyAI operator

creates

configure

creates

Project 1

ModelMesh Serving Runtime

Service Monitor

Route

TrustyAI Service 1

PVC-1

Project 2

ModelMesh Serving Runtime

Service Monitor

Route

TrustyAI Service 2

PVC-2

PV

TrustyAI external library integration

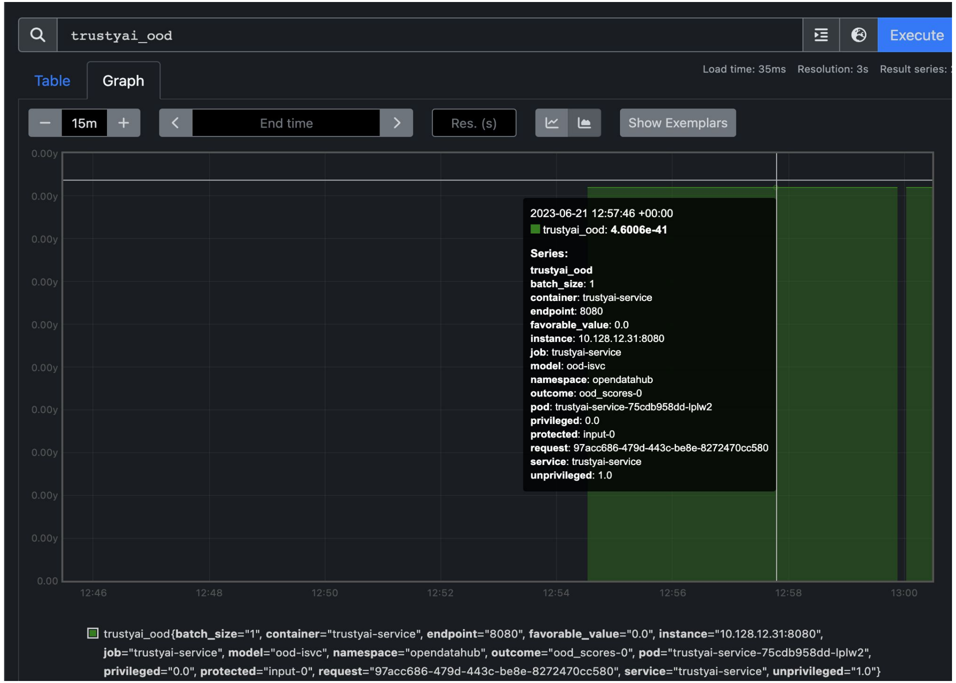
External library support

- TrustyAI service
  - Inclusion of AIX360’s TSICE Explainer
  - Explainer endpoints
  - Exclusion of synthetic metadata
  - Time-series support in TrustyAI dataframes
  - `TSSaliency`
- Python
  - Addition of external algorithms
    - `pip install trustyai[extras]`
    - TSICE explainer

OOD Models

OOD Model support

- Works *mostly* out-of-the-box
- TrustyAI service
    - ○ Create a Passthrough metric that takes a model output and registers it as a Prometheus metric
    - ○ Optional aggregations of outputs?
        - ■ Mean, max, min, etc?



```
trustyai_ood
```

```
2023-06-21 12:57:46 +00:00
trustyai_ood: 4.6006e-41

Series:
trustyai_ood
batch_size: 1
container: trustyai-service
endpoint: 8080
favorable_value: 0.0
instance: 10.128.12.31:8080
job: trustyai-service
model: ood-isvc
namespace: opendatahub
outcome: ood_scores-0
pod: trustyai-service-75cdb958dd-lplw2
privileged: 0.0
protected: input-0
request: 97acc686-479d-443c-be8e-8272470cc580
service: trustyai-service
unprivileged: 1.0
```

🟩 trustyai_ood{**batch_size**="1", **container**="trustyai-service", **endpoint**="8080", **favorable_value**="0.0", **instance**="10.128.12.31:8080", **job**="trustyai-service", **model**="ood-isvc", **namespace**="opendatahub", **outcome**="ood_scores-0", **pod**="trustyai-service-75cdb958dd-lplw2", **privileged**="0.0", **protected**="input-0", **request**="97acc686-479d-443c-be8e-8272470cc580", **service**="trustyai-service", **unprivileged**="1.0"}

Community

Community

- GitHub Projects[1]
    - Experimental planning board
        - https://github.com/orgs/trustyai-explainability/projects/12
    - Roadmap
        - https://github.com/orgs/trustyai-explainability/projects/10

[1]: https://github.com/orgs/trustyai-explainability/projects

Roadmap

TrustyAI 2023 roadmap

July 2023 TrustyAI 0.2.0

- *Explainers*
  - Support for explainers LIME, SHAP, CF at service level
- *Metrics*
  - Flexible scheduling/batching
  - Improve service metadata endpoints
    - Include available categories
- *Operator*
  - TrustyAI Operator v1

September 2023 TrustyAI 0.3.0

- *Explainers*
  - Support for external explainability libraries
- *Metrics*
  - Additional metrics
  - Metrics statistical tests

December 2023 TrustyAI 0.4.0

- *Storage*
  - Wider storage support (database backends)
- *Explainers*
  - NLP explainability support
- *Metrics*
  - Support for user-defined historical windows