

Odahu Platform Tutorial

version 1.0

Odahu Team

February 04, 2020

Contents

| | |
|--|----|
| About Open Data Analytics Hub (ODAHU) | 1 |
| Concepts | 3 |
| Phases | 3 |
| Toolchains | 3 |
| Ready to use | 3 |
| Model storage | 4 |
| Architecture | 5 |
| Distributions | 6 |
| HELM charts | 6 |
| Docker Images | 6 |
| Python packages | 7 |
| NPM packages | 7 |
| Security | 9 |
| Securing model APIs | 9 |
| Securing management APIs | 9 |
| Securing network connections, ingress and egress | 9 |
| Installation | 10 |
| Kubernetes setup | 10 |
| Deploy Kubernetes cluster in Google Compute Platform (GKE) | 10 |
| Deploy Kubernetes cluster in Amazon Web Services (EKS) | 11 |
| Deploy Kubernetes cluster in Microsoft Azure (AKS) | 11 |
| Install base Kubernetes services | 12 |
| Install Helm and Tiller (version 2.14.3) | 12 |
| Install Nginx Ingress | 12 |
| Install Istio (with Helm and Tiller) | 13 |
| Add ODAHU Helm charts repository | 14 |
| Install Knative | 14 |
| Install Tekton Pipelines | 14 |
| Install Fluentd with set of cloud object storage plugins | 14 |
| Install ODAHU | 15 |
| Install core ODAHU services | 15 |
| Training service (MLFlow) | 17 |
| Packaging service | 17 |
| Install additional services (optional) | 17 |
| Delete ODAHU services | 17 |

| | |
|---|----|
| Conclusion | 18 |
| Quickstart | 19 |
| Prerequisites | 19 |
| Tutorial | 19 |
| Create MLFlow project | 19 |
| Setup connections | 22 |
| Create a Connection to GitHub repository | 23 |
| Create Connection to wine-quality.csv object storage | 23 |
| Create a Connection to a docker registry | 24 |
| Train the model | 25 |
| Package the model | 27 |
| Deploy the model | 28 |
| Use the deployed model | 29 |
| API | 31 |
| API-provided URLs | 31 |
| Authentication and authorization | 31 |
| Implementation details | 31 |
| Operator | 32 |
| Implementation details | 32 |
| MLFlow Trainer | 33 |
| Limitations | 33 |
| Implementation Details | 33 |
| Metrics | 34 |
| Airflow | 35 |
| Connections | 35 |
| Odahu-flow Connection | 35 |
| Configuring the Connection | 35 |
| Custom operators | 36 |
| Train, Pack, Deploy operators | 36 |
| Model usage operators | 37 |
| Helper operators | 38 |
| How to describe operators | 38 |
| DAG example | 39 |
| JupyterLab extension | 43 |
| Installation | 43 |
| Configuration | 43 |
| Login | 44 |

| | |
|--|----|
| Usage | 44 |
| Templates | 44 |
| Main view | 45 |
| Log viewer | 46 |
| Submit resources | 47 |
| Odahuflowctl | 48 |
| Installation | 48 |
| Help | 48 |
| Login | 48 |
| Specifying of a token explicitly | 48 |
| Sign in interactively | 48 |
| Model Format | 50 |
| odahuflow.model.yaml | 50 |
| Odahu Model Environments | 51 |
| Odahu's General Python Prediction Interface (GPPI) | 51 |
| General Information | 51 |
| Description | 51 |
| Required Environment variables | 51 |
| Interface declaration | 52 |
| Connections | 53 |
| General connection structure | 53 |
| Connection management | 53 |
| Swagger UI | 53 |
| Odahu-flow CLI | 53 |
| JupyterLab | 54 |
| Connection types | 54 |
| S3 | 54 |
| Google Cloud Storage | 55 |
| Azure Blob storage | 56 |
| GIT | 57 |
| Docker | 57 |
| Amazon Elastic Container Registry | 58 |
| Model Trainings | 60 |
| General training structure | 60 |
| Training data | 61 |
| Trainings management | 61 |
| Swagger UI | 62 |

| | |
|--|----|
| Odahu-flow CLI | 62 |
| JupyterLab | 62 |
| MLFlow | 62 |
| Installation | 62 |
| MLProject file | 63 |
| MLFlow protocol | 63 |
| Model Packagers | 65 |
| Installation | 65 |
| General packager structure | 65 |
| Packagers management | 66 |
| Swagger UI | 66 |
| Odahu-flow CLI | 66 |
| JupyterLab | 67 |
| Docker REST | 67 |
| Docker CLI | 69 |
| Model Deployments | 74 |
| General deployment structure | 74 |
| Model Deployment management | 74 |
| Swagger UI | 74 |
| Odahu-flow CLI | 74 |
| JupyterLab | 75 |
| Service Catalog | 75 |
| Grafana Dashboard | 76 |
| Feedback | 77 |
| Protocol | 78 |
| Working example | 78 |
| Working Example - Send Feedback as Payload | 79 |
| Glossary | 80 |
| Development | 83 |
| Pre requirements | 83 |
| Useful links | 83 |
| A repository directory structure | 85 |
| Repositories | 86 |
| odahu/odahu-flow | 86 |
| odahu/odahu-trainer | 86 |
| odahu/odahu-packager | 86 |
| odahu/odahu-flow-jupyterlab-plugin | 86 |

| | |
|----------------------------------|----|
| odahu/odahu-airflow-plugin | 86 |
| odahu/odahu-docs | 86 |
| odahu/odahu-examples | 86 |
| odahu/odahu-infra | 88 |
| Development hints | 89 |
| Set up a development environment | 89 |
| Update dependencies | 90 |
| Make changes in API entities | 90 |
| Actions before a pull request | 90 |
| Local Helm deploy | 90 |
| Integration Testing | 92 |
| Running system/integration tests | 92 |
| Indices and tables | 94 |
| Index | 95 |

About Open Data Analytics Hub (ODAHU)

The ODAHU project is focused on creating system of services, extensions for third party systems and tools which help to accelerate building enterprise level systems with automated AI/ML models life cycle.

Multi ML Frameworks

- Supporting major ML frameworks: Scikit-learn, Keras, Tensorflow, PyTorch, H2O (and more)
- Extends MLflow platform with enterprise level features

Multi Infrastructure

- Kubernetes native system of services
- Deployment automation to Kubernetes cluster with Helm charts
- Supporting major Kubernetes platforms: AWS EKS, Azure AKS, GCP GKE, RedHat OpenShift

Secure

- Single sign-on (SSO) based on OAuth2
- RESTful API secured with SSL
- Role based access control
- Users activity audit
- Credentials and keys manager based on HashiCorp Vault
- Internal traffic encryption with Istio

Modular and Extensible

- Services for different ML phases: transform, train, package, deploy, evaluate
- Services are extensible and manageable via REST APIs, SDK and CLI
- Functionality can be extended with new services
- Connectors for data sources
- Connectors for package repositories
- Connectors for Docker container registries
- Plugins for data science IDEs
- Plugins for workflow engines (like Airflow and others)

Scalable

- Systems based on ODAHU can be scaled from small to very large.
- Scalable ML model training phase
- Scalable ML model serving phase

Manageable

- Pre-build monitoring dashboards
- Configurable alerting rules

About Open Data Analytics Hub (ODAHU)

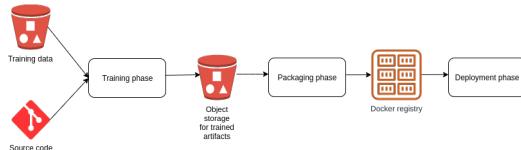
- Configurable logs collection
- Compatible with third party logs processing system

Open

- It is free and open-source with the Apache2 License.
- Contribution to project is welcome!

Concepts

Phases



Odahu splits the ML/AI model lifecycle into three phases:

1. **Train**
2. **Package**
3. **Deploy**

Applications and tools can further automate each phase by implementing pluggable extensions as

1. **Trainer**
2. **Packager** or
3. **Deployer**

Trainers and Packagers can be registered as components of the Odahu Platform using:

1. **Trainer Extension**
2. **Packager Extension**

When registered, these components can use Odahu **Trainer Metrics** and **Trainer Tags**.

Users are encouraged to integrate third-party **Trainer Extensions** and **Packager Extensions**.

Toolchains

Taken together a Trainer, Packager, and Deployer comprise a *Toolchain* that automates an end-to-end machine learning pipeline.

Ready to use

Odahu provides a **Trainer Extension** and a **Packager Extension**

1. **MLflow Trainer**
2. **REST API Packager**

These power the default Toolchain.

Model storage

Odahu Platform stores models in **Trained Model Binaries** for different languages.

Presently, Odahu Platform supports only:

1. General Python Prediction Interface

Users are encouraged to provide additional formats.

Architecture

The following diagram shows the high level architecture of Odahu-flow project.



Odahu-flow applications:

- Odahu-flow CLI is a command-line interface for interacting with Odahu-flow API service.
- The Odahu-flow Swagger UI provides swagger interface for Odahu-flow REST API.
- Odahu-flow Airflow plugin provides a set of custom operators that allow you to interact with a Odahu cluster using [Apache Airflow](#)
- The MLflow Tracking component is an API and UI for logging parameters, code versions, and metrics when running your machine learning code and for later visualizing the results.
- JupyterLab extension allows you to interact with an Odahu cluster from JupyterLab web-based IDEs.
- API service manages Odahu Platform entities: Connections, Trainings, Packaging, Deployments.
- Service catalog provides a Swagger UI for Model Deployments.
- TektonCD is an open source implementation to configure and run CI/CD style pipelines for your Kubernetes application.
- Knative Serving builds on Kubernetes and Istio to support deploying and serving of serverless applications and functions.
- Odahu-flow Model Training API provides features to manage remote training jobs.
- Odahu-flow packagers turn a **Trained Model Binary** artifact into a specific application.
- Odahu-flow Model Deployment API allows deploy ML models in a Kubernetes cluster.

Distributions

HELM charts

- Release and pre-release **Helm charts** are in github.

| Helm chart name | Repository | Description |
|------------------------|----------------------|--------------------------------------|
| odahu-flow-fluentd | odahu/odahu-infra | Fluentd with gcp, s3 and abs plugins |
| odahu-flow-k8s-gke-saa | odahu/odahu-infra | GKE role assigner |
| odahu-flow-knative | odahu/odahu-infra | Custom knative chart |
| odahu-flow-monitoring | odahu/odahu-infra | Prometheus, grafana and alertmanager |
| odahu-flow-tekton | odahu/odahu-infra | Custom tekton chart |
| odahu-flow-core | odahu/odahu-flow | Core Odahu-flow services |
| odahu-flow-mlflow | odahu/odahu-trainer | Odahu-flow mlflow toolchain |
| odahu-flow-packagers | odahu/odahu-packager | Odahu-flow REST packager |

Docker Images

Release versions of images are on Docker Hub in the [odahu](#) team.

| Image name | Repository | Description |
|----------------------------|-------------------|--------------------------------------|
| odahu-flow-fluentd | odahu/odahu-infra | Fluentd with gcp, s3 and abs plugins |
| odahu-flow-api | odahu/odahu-flow | Odahu-flow API service |
| odahu-flow-model-cli | odahu/odahu-flow | Odahu-flow CLI |
| odahu-flow-model-trainer | odahu/odahu-flow | Trainer helper |
| odahu-flow-model-packager | odahu/odahu-flow | Packager helper |
| odahu-flow-service-catalog | odahu/odahu-flow | Swagger for model deployments |

Python packages

| | | |
|-----------------------------------|------------------------------------|---|
| odahu-flow-operator | odahu/odahu-flow | Odahu-flow kubernetes orchestrator |
| odahu-flow-feedback-collector | odahu/odahu-flow | REST API for user feedback service |
| odahu-flow-feedback-rq-catcher | odahu/odahu-flow | Model deployment request-response catcher |
| odahu-flow-mlflow-toolchain | odahu/odahu-trainer | Odahu-flow mlflow toolchain |
| odahu-flow-mlflow-toolchain-gpu | odahu/odahu-trainer | Odahu-flow mlflow toolchain with NVIDIA GPU |
| odahu-flow-mlflow-tracking-server | odahu/odahu-trainer | MLflow tracking service |
| odahu-flow-packagers | odahu/odahu-packager | Odahu-flow packagers |
| base-notebook | odahu/odahu-flow-jupyterlab-plugin | Image with JupyterLab extension based on jupyter/base-notebook |
| datascience-notebook | odahu/odahu-flow-jupyterlab-plugin | Image with JupyterLab extension based on jupyter/datascience-notebook |
| tensorflow-notebook | odahu/odahu-flow-jupyterlab-plugin | Image with JupyterLab extension based on jupyter/tensorflow-notebook |

Python packages

- Release versions of Python packages are on PyPi in project odahu.

| Package name | Repository | Description |
|------------------------------|------------------------------------|---|
| odahu-flow-cli | odahu/odahu-flow | Odahu-flow CLI |
| odahu-flow-sdk | odahu/odahu-flow | Odahu-flow SDK |
| odahu-flow-jupyterlab-plugin | odahu/odahu-flow-jupyterlab-plugin | Jupyterlab with the Odahu-flow plugin |
| odahu-flow-airflow-plugin | odahu/odahu-airflow-plugin | Odahu-flow Airflow plugin(operators, hooks and so on) |

NPM packages

- Release versions of Python packages are on npm in project odahu.

| Package name | Repository | Description |
|--------------|------------|-------------|
|--------------|------------|-------------|

Python packages

| | | |
|------------------------------|------------------------------------|---------------------------------------|
| odahu-flow-jupyterlab-plugin | odahu/odahu-flow-jupyterlab-plugin | Jupyterlab with the Odahu-flow plugin |
|------------------------------|------------------------------------|---------------------------------------|

Security

Odahu delegates user, token and credential management to third-party providers. Any service that supports the OAuth2 protocol can be used to provide authentication and authorization to a Odahu cluster.

Warning

Connectivity from cluster to OAuth2 IP (identity provider) is required.

Odahu integrates OAuth2 to Identity Providers using the `oauth2_proxy` library.

Securing model APIs

Model prediction API and **Model Feedback API** calls are secured using **JWT tokens**. **API service** can issue a new **JWT token** for a given **role name**. **JWT tokens** are scoped to specific **role names** and may be used for granting access to the **Model prediction API**. The role name has to be defined during creation of **model deployments**, otherwise a hard-coded default value will be used.

Securing management APIs

When enabled in the Odahu start-up configuration, all other endpoints (such as **API service**) are secured using **JWT tokens** and OAuth2 cookies. When a user tries to open a secured resource, `oauth2_proxy` checks the incoming request for a **JWT token** and OAuth2 Cookies. If a request does not contain appropriate credentials, the user will be redirected to a login URL.

Securing network connections, ingress and egress

Odahu does not ship with network policies enabled, but any policy that uses Pod label selection can be used (e.g. K8s NetworkPolicy). Details are in the Network Connectivity Reference.

Installation

To install ODAHU services, you need to provide a number of preliminary requirements for it.

In particular:

- Kubernetes cluster to perform base and accessory ODAHU services in it, as well as models training, packaging and deployment processes. To be able to use ODAHU services, minimum version of your Kubernetes cluster must be at least [1.13](#).
- object storage to store models training artifacts and get input data for models (S3, Google Cloud Storage, Azure Blob storage are supported)
- Docker registry (to store resulting Docker images from packagers)

Kubernetes setup

Deploy Kubernetes cluster in Google Compute Platform (GKE)

Prerequisites:

- [GCP service account](#) to deploy Kubernetes cluster with and use it credentials for access to object storage and Google Cloud Registry
- Google Cloud Storage bucket (odahu-flow-test-store in examples below) to store models output data

Run deploy of a new Kubernetes cluster:

```
$ gcloud container clusters create <cluster-name> \
--cluster-version 1.13 \
--machine-type=n1-standard-2 \
--disk-size=100GB \
--disk-type=pd-ssd \
--num-nodes 4 \
--zone <cluster-region> \
--project <project-id>
```

Note

Make sure that the disk size on the cluster nodes is sufficient to store images for all services and packaged models. We recommend using a disk size of at least 100 GiB.

Fetch your Kubernetes credentials for kubectl after cluster is successfully deployed:

```
$ gcloud container clusters get-credentials <cluster-name> \
--zone <cluster-region> \
--project <project-id>
```

Deploy Kubernetes cluster in Amazon Web Services (EKS)

Prerequisites

- Resources that are [described in AWS documentation](#)
- AWS S3 bucket (odahu-flow-test-store in examples below) to store models output data

After you've created VPC and a dedicated security group for it along with Amazon EKS service role to apply to your cluster, you can create a Kubernetes cluster with following command:

```
$ aws eks --region <cluster-region> create-cluster \  
--name <cluster-name> --kubernetes-version 1.13 \  
--role-arn arn:aws:iam::111122223333:role/eks-service-role-AWSServiceRoleForAmazonEKS-EXAMPLEBKZRQR \  
--resources-vpc-config subnetIds=subnet-a9189fe2,subnet-50432629,securityGroupIds=sg-f5c54184
```

Use the AWS CLI update-kubeconfig command to create or update kubeconfig for your cluster:

```
$ aws eks --region <cluster-region> update-kubeconfig --name <cluster-name>
```

Deploy Kubernetes cluster in Microsoft Azure (AKS)

Prerequisites

- [Azure AD Service Principal](#) to interact with Azure APIs and create dynamic resources for an AKS cluster
- Azure Storage account with Blob container (odahu-flow-test-store in examples below) to store models output data

First, create a resource group in which all created resources will be placed:

```
$ az group create --location <cluster-region> \  
--name <resource-group-name>
```

Run deploy of a new Kubernetes cluster:

```
$ az aks create --name <cluster-name> \  
--resource-group <resource-group-name> \  
--node-vm-size Standard_DS2_v2 \  
--node-osdisk-size 100GB \  
--node-count 4 \  
--service-principal <service-principal-appid> \  
--client-secret <service-principal-password>
```

Fetch your Kubernetes credentials for kubectl after cluster is successfully deployed:

```
$ az aks get-credentials --name <cluster-name> \  
--resource-group <resource-group-name>
```

Install base Kubernetes services

Install base Kubernetes services

Install Helm and Tiller (version 2.14.3)

Make sure you have a Kubernetes service account with the cluster-admin role defined for Tiller. If not already defined, create one using following commands:

```
$ cat << EOF > tiller_sa.yaml
---
apiVersion: v1
kind: ServiceAccount
metadata:
  name: tiller
  namespace: kube-system
---
apiVersion: rbac.authorization.k8s.io/v1
kind: ClusterRoleBinding
metadata:
  name: tiller
roleRef:
  apiGroup: rbac.authorization.k8s.io
  kind: ClusterRole
  name: cluster-admin
subjects:
- kind: ServiceAccount
  name: tiller
  namespace: kube-system
EOF
```

```
$ kubectl apply -f ./tiller_sa.yaml
```

Install Tiller in your cluster with created service account:

```
$ helm init --service-account=tiller
```

Ensure that Tiller is installed:

```
$ kubectl -n kube-system get pods --selector=app=helm
NAME           READY   STATUS    RESTARTS   AGE
tiller-deploy-57f498469-r5cmv   1/1     Running   0          16s
```

Install Nginx Ingress

Install nginx-ingress Helm chart:

```
$ helm install stable/nginx-ingress --name nginx-ingress --namespace kube-system
```

Get external LoadBalancer IP assigned to nginx-ingress service:

```
$ kubectl get -n kube-system svc nginx-ingress-controller \
-o=jsonpath='{.status.loadBalancer.ingress[*].ip}{"\n"}'
```

Install Istio (with Helm and Tiller)

Note

ODAHU services uses number of Istio custom resources actively, so Istio installation is mandatory.

Add Helm repository for Istio charts

```
$ helm repo add istio https://storage.googleapis.com/istio-release/releases/1.4.2/charts/
```

Create a namespace for the istio-system components

```
$ kubectl create namespace istio-system
```

Install the `istio-init` chart to bootstrap all the Istio's CustomResourceDefinitions

```
$ helm install istio/istio-init --name istio-init --namespace istio-system
```

Ensure that all `istio-init` jobs have been completed:

```
$ kubectl -n istio-system get job \
-o=jsonpath='{range.items[?(@.status.succeeded==1)]}{.metadata.name}{"\n"}{end}'
```

Install Istio Helm chart with provided values.

Example:

```
$ cat << EOF > istio_values.yaml
global:
  proxy:
    accessLogFile: "/dev/stdout"
    disablePolicyChecks: false
  sidecarInjectorWebhook:
    enabled: true
  pilot:
    enabled: true
  mixer:
    policy:
      enabled: true
    telemetry:
      enabled: true
  adapters:
    stdio:
      enabled: true
  gateways:
    istio-ingressgateway:
      enabled: true
      type: ClusterIP
      meshExpansionPorts: []
      ports:
        - port: 80
          targetPort: 80
          name: http
        - port: 443
          name: https
        - port: 15000
EOF
```

Install base Kubernetes services

```
name: administration
istio-egressgateway:
  enabled: true
prometheus:
  enabled: false
EOF

$ helm install istio/istio --name istio --namespace istio-system --values ./istio_values.yaml
```

Add ODAHU Helm charts repository

```
$ helm repo add odahu https://raw.githubusercontent.com/odahu/odahu-helm/master
```

Install Knative

Create namespace for Knative and label it for Istio injection:

```
$ kubectl create namespace knative-serving && \
  kubectl label namespace knative-serving istio-injection=enabled
```

Install Knative with [Helm chart](#) provided by ODAHU team:

```
$ helm install odahu/odahu-flow-knative --name knative --namespace knative-serving
```

Install Tekton Pipelines

Create namespace for Tekton:

```
$ kubectl create namespace tekton-pipelines
```

Install Tekton Pipelines with [Helm chart](#) provided by ODAHU team:

```
$ helm install odahu/odahu-flow-tekton --name tekton --namespace tekton-pipelines
```

Install Fluentd with set of cloud object storage plugins

In order to save models training logs to object storage of cloud provider you use, a container with fluentd is used, in which a set of [plugins](#) for popular cloud providers' object storages (AWS S3, Google Storage, Azure Blob) is added. Installation is done using a [fluentd Helm chart](#) provided by ODAHU team.

First, create an object storage bucket:

Install ODAHU

```
$ gsutil mb gs://odahu-flow-test-store/
```

Create namespace for Fluentd:

```
$ kubectl create namespace fluentd
```

Install Fluentd with specified values. Full list of values you can see in chart's [values.yaml](#).

Example:

```
$ cat << EOF > fluentd_values.yaml
output:
  target: gcs
  gcs:
    authorization: keyfile
    bucket: odahu-flow-test-store
    project: my-gcp-project-id-zzzzz
    private_key_id: 0bacc0b0caa0a0aacabacbab0a0b00ababacaab
    private_key: -----BEGIN PRIVATE KEY-----\nprivate-key-here\n-----END PRIVATE KEY-----\n
    client_email: service-account@my-gcp-project-id-zzzzz.iam.gserviceaccount.com
    client_id: 00000000000000000000
    auth_uri: https://accounts.google.com/o/oauth2/auth
    token_uri: https://oauth2.googleapis.com/token
    auth_provider_x509_cert_url: https://www.googleapis.com/oauth2/v1/certs
    client_x509_cert_url: https://www.googleapis.com/robot/v1/metadata/x509/service-account%40my-gcp-project-id-zzzzz.iam.gserviceaccount.com
EOF
```

```
$ helm install odahu/odahu-flow-fluentd --name fluentd --namespace fluentd --values ./fluentd_values.yaml
```

Install ODAHU

Install core ODAHU services

Create namespace for core ODAHU service:

```
$ kubectl create namespace odahu-flow &&
  kubectl label namespace odahu-flow project=odahu-flow
```

Create namespaces for ODAHU training, packaging and deployment.

```
$ for i in training packaging deployment; do \
  kubectl create namespace odahu-flow-$i &&
  kubectl label namespace odahu-flow-$i project=odahu-flow; done
```

Deployment namespace should be also labeled for Istio injection.

```
$ kubectl label namespace odahu-flow-deployment istio-injection=enabled
```

Prepare YAML config with values for `odahu-flow-core` Helm chart.

Example:

Install ODAHU

```
$ cat << EOF > odahuflow_values.yaml
logLevel: debug
ingress:
  enabled: true
  globalDomain: odahu.example.com
edge:
  ingress:
    enabled: true
    domain: odahu.example.com
feedback:
  enabled: true
config:
  common:
    external_urls:
      - name: Documentation
        url: https://docs.odahu.org
connection:
  enabled: true
  decrypt_token: somenotemptystring
  repository_type: kubernetes
deployment:
  edge:
    host: http://odahu.example.com
EOF
```

Note

This example uses hostname `odahu.example.com` as entrypoint for cluster services. It should point to LoadBalancer IP got from Nginx Ingress section.

In order to setup ODAHU services along with ready-to-use **connections**, you may add according section to values YAML in advance.

Examples:

- Docker registry connection is used to pull/push Odahu packager resulting Docker images to a Docker registry

```
connections:
- id: docker-hub
  spec:
    description: Docker registry for model packaging
    username: "user"
    password: "supersecure"
    type: docker
    uri: docker.io/odahu-models-repo
    webUILink: https://hub.docker.com/r/odahu-models-repo
```

- Google Cloud Storage connection is used to store model trained artifacts and

Install `odahu-flow` core services:

```
$ helm install odahu/odahu-flow-core --name odahu-flow --namespace odahu-flow --values ./odahuflow_values.yaml
```

Delete ODAHU services

Training service (*MLFlow*)

Prepare YAML config with values for `odahu-flow-mlflow` Helm chart.

```
$ cat << EOF > mlflow_values.yaml
logLevel: debug
ingress:
  globalDomain: example.com
  enabled: true
tracking_server:
  annotations:
    sidecar.istio.io/inject: "false"
toolchain_integration:
  enabled: true
EOF
```

Install Helm chart:

```
$ helm install odahu/odahu-flow-mlflow --name odahu-flow-mlflow --namespace odahu-flow \
--values ./mlflow_values.yaml
```

Packaging service

Install `odahu-flow-packagers` Helm chart:

```
$ helm install odahu/odahu-flow-packagers --name odahu-flow-packagers --namespace odahu-flow
```

Install additional services (optional)

In order to provide additional functionality, ODAHU team also developed several Helm charts to install them into Kubernetes cluster. These are:

- `odahu-flow-monitoring` - Helm chart providing installation and setup of
 - `Prometheus operator` - to collect various metrics from models trainings
 - `Grafana` with set of custom dashboards - to visualize these metrics
- `odahu-flow-k8s-gke-saa` - Helm chart providing installation and setup of `k8s-gke-service-account-assigner` service.

Delete ODAHU services

To delete and purge Helm chart run:

```
$ helm delete --purge odahu-flow
```

To clean up remaining CustomResourceDefinitions execute following command:

Conclusion

```
$ kubectl get crd | awk '/odahuflow/ {print $1}' | xargs -n1 kubectl delete crd
```

To purge everything installed in previous steps with single command, run

```
$ helm delete --purge odahu-flow-packagers odahu-flow-mlflow odahu-flow &&\\
kubectl delete namespace odahu-flow &&\\
for i in training packaging deployment; do \\
    kubectl delete namespace odahu-flow-${i} || true; done &&\\
kubectl get crd | awk '/odahuflow/ {print $1}' | xargs -n1 kubectl delete crd &&\\
kubectl -n istio-system delete job.batch/odahu-flow-feedback-rq-catcher-patcher &&\\
kubectl -n istio-system delete sa/odahu-flow-feedback-rq-catcher-patcher &&\\
kubectl -n istio-system delete cm/odahu-flow-feedback-rq-catcher-patch
```

Conclusion

After successful deployment of a cluster, you may proceed to Quickstart section and learn how to perform base ML operations such as **train**, **package** and **deploy** steps.

Quickstart

In this tutorial you will learn how to Train, Package and Deploy a model from scratch on Odahu. Once deployed, the model serves RESTful requests, and makes a prediction when provided user input.

Odahu's API server performs Train, Package, and Deploy operations for you, using its REST API.

Prerequisites

- Odahu cluster
- MLFlow and **REST API Packager** (installed by default)
- **Odahu-flow CLI** or **Plugin for JupyterLab** (installation instructions: CLI, Plugin)
- JWT token from API (instructions)
- Google Cloud Storage bucket on Google Compute Platform
- GitHub repository and an ssh key to connect to it

Tutorial

In this tutorial, you will learn how to:

1. Create an MLFlow project
2. Setup Connections
3. Train a model
4. Package the model
5. Deploy the packaged model
6. Use the deployed model

This tutorial uses a dataset to predict the quality of the wine based on quantitative features like the wine's *fixed acidity*, *pH*, *residual sugar*, and so on. The dataset is from [UCI's machine learning repository](#).

Code for the tutorial is available on [GitHub](#).

Create MLFlow project

| | |
|--------|--|
| Before | Odahu cluster that meets prerequisites |
| After | Model code that predicts wine quality |

Create a new project folder:

Quickstart

```
$ mkdir wine && cd wine
```

Create a training script:

```
$ touch train.py
```

Paste code into the file:

train.py

```
1 import os
2 import warnings
3 import sys
4 import argparse
5
6 import pandas as pd
7 import numpy as np
8 from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
9 from sklearn.model_selection import train_test_split
10 from sklearn.linear_model import ElasticNet
11
12 import mlflow
13 import mlflow.sklearn
14
15 def eval_metrics(actual, pred):
16     rmse = np.sqrt(mean_squared_error(actual, pred))
17     mae = mean_absolute_error(actual, pred)
18     r2 = r2_score(actual, pred)
19     return rmse, mae, r2
20
21 if __name__ == "__main__":
22     warnings.filterwarnings("ignore")
23     np.random.seed(40)
24
25     parser = argparse.ArgumentParser()
26     parser.add_argument('--alpha')
27     parser.add_argument('--l1-ratio')
28     args = parser.parse_args()
29
30     # Read the wine-quality csv file (make sure you're running this from the root of MLflow!)
31     wine_path = os.path.join(os.path.dirname(os.path.abspath(__file__)), "wine-quality.csv")
32     data = pd.read_csv(wine_path)
33
34     # Split the data into training and test sets. (0.75, 0.25) split.
35     train, test = train_test_split(data)
36
37     # The predicted column is "quality" which is a scalar from [3, 9]
38     train_x = train.drop(["quality"], axis=1)
39     test_x = test.drop(["quality"], axis=1)
40     train_y = train[["quality"]]
41     test_y = test[["quality"]]
42
43     alpha = float(args.alpha)
44     l1_ratio = float(args.l1_ratio)
45
46     with mlflow.start_run():
47         lr = ElasticNet(alpha=alpha, l1_ratio=l1_ratio, random_state=42)
48         lr.fit(train_x, train_y)
49
50         predicted_qualities = lr.predict(test_x)
51
52         (rmse, mae, r2) = eval_metrics(test_y, predicted_qualities)
53
54         print("Elasticnet model (alpha=%f, l1_ratio=%f):" % (alpha, l1_ratio))
55         print("  RMSE: %s" % rmse)
56         print("  MAE: %s" % mae)
57         print("  R2: %s" % r2)
```

Quickstart

```
58
59     mlflow.log_param("alpha", alpha)
60     mlflow.log_param("l1_ratio", l1_ratio)
61     mlflow.log_metric("rmse", rmse)
62     mlflow.log_metric("r2", r2)
63     mlflow.log_metric("mae", mae)
64     mlflow.set_tag("test", '13')
65
66     mlflow.sklearn.log_model(lr, "model")
67
68 # Persist samples (input and output)
69 train_x.head().to_pickle('head_input.pkl')
70 mlflow.log_artifact('head_input.pkl', 'model')
71 train_y.head().to_pickle('head_output.pkl')
72 mlflow.log_artifact('head_output.pkl', 'model')
```

In this file, we:

- Start MLflow context on line 46
- Train ElasticNet model on line 48
- Set metrics, parameters and tags on lines 59-64
- Save model with name `model` (`model` is serialized and sent to the MLflow engine) on line 66
- Save input and output samples (for persisting information about input and output column names) on lines 69-72

Create an `MLproject` file:

```
$ touch MLproject
```

Paste code into the file:

MLproject

```
name: wine-quality-example
conda_env: conda.yaml
entry_points:
  main:
    parameters:
      alpha: float
      l1_ratio: {type: float, default: 0.1}
    command: "python train.py --alpha {alpha} --l1-ratio {l1_ratio}"
```

Note

Read more about `MLproject` structure on the [official MLFlow docs](#).

Create a `conda` environment file:

```
$ touch conda.yaml
```

Paste code to the created file:

conda.yaml

Quickstart

```
name: example
channels:
- defaults
dependencies:
- python=3.6
- numpy=1.14.3
- pandas=0.22.0
- scikit-learn=0.19.1
- pip:
- mlflow==1.0.0
```

Note

All python packages that are used in training script must be listed in the conda.yaml file.

Read more about conda environment on the [official conda docs](#).

Make directory “data” and download the wine data set:

```
$ mkdir ./data
$ wget https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv -O ./data/wine-quality.csv
```

After this step the project folder should look like this:

```
.
├── MLproject
├── conda.yaml
├── data
│   └── wine-quality.csv
└── train.py
```

Setup connections

| | |
|--------|--|
| Before | Odahu cluster that meets prerequisites |
| After | Odahu cluster with Connections |

Odahu Platform uses the concept of **Connections** to manage authorizations to external services and data.

This tutorial requires three Connections:

- A GitHub repository, where the code is located
- A Google Cloud Storage folder, where input data is located (wine-quality.csv)
- A Docker registry, where the trained and packaged model will be stored for later use

You can find more detailed documentation about a connection configuration [here](#).

Create a Connection to GitHub repository

Because [odahu-examples](#) repository already contains the required code we will just use this repository. But feel free to create and use a new repository if you want.

Odahu is REST-powered, and so we encode the REST “payloads” in this tutorial in YAML files. Create a directory where payloads files will be staged:

```
$ mkdir ./odahu-flow
```

Create payload:

```
$ touch ./odahu-flow/vcs_connection.odahu.yaml
```

Paste code into the created file:

vcs_connection.odahu.yaml

```
kind: Connection
id: odahu-flow-tutorial
spec:
  type: git
  uri: git@github.com:odahu/odahu-examples.git
  reference: origin/master
  keySecret: <paste here your key github ssh key>
  description: Git repository with odahu-flow-examples
  webUILink: https://github.com/odahu/odahu-examples
```

Note

Read more about [GitHub ssh keys](#)

Create a Connection using the **Odahu-flow CLI**:

```
$ odahuflowctl conn create -f ./odahu-flow/vcs_connection.odahu.yaml
```

Or create a Connection using **Plugin for JupyterLab**:

1. Open jupyterlab (available by <your.cluster.base.address>/jupyterhub);
2. Navigate to ‘File Browser’ (folder icon)
3. Select file `./odahu-flow/vcs_connection.odahu.yaml` and in context menu press submit button;

Create Connection to wine-quality.csv object storage

Create payload:

```
$ touch ./odahu-flow/wine_connection.odahu.yaml
```

Quickstart

Paste this code into the file:

wine_connection.odahu.yaml

```
kind: Connection
id: wine-tutorial
spec:
  type: gcs
  uri: gs://<paste your bucket address here>/data-tutorial/wine-quality.csv
  region: <paste region here>
  keySecret: <paste key secret here> # should be enclosed in single quotes
  description: Wine dataset
```

Create a connection using the **Odahu-flow CLI** or **Plugin for JupyterLab**, as in the previous example.

If wine-quality.csv is not in the GCS bucket yet, use this command:

```
$ gsutil cp ./data/wine-quality.csv gs://<bucket-name>/data-tutorial/
```

Create a Connection to a docker registry

Create payload:

```
$ touch ./odahu-flow/docker_connection.odahu.yaml
```

Paste this code into the file:

docker_connection.odahu.yaml

```
kind: Connection # type of payload
id: docker-tutorial
spec:
  type: docker
  uri: <paste uri of your registry here> # uri to docker image registry
  username: <paste your username here>
  password: <paste your password here>
  description: Docker registry for model packaging
```

Create the connection using **Odahu-flow CLI** or **Plugin for JupyterLab**, as in the previous example.

Check that all Connections were created successfully:

```
- id: docker-tutorial
  description: Docker repository for model packaging
  type: docker
- id: odahu-flow-tutorial
  description: Git repository with odahu-flow-tutorial
  type: git
- id: models-output
  description: Storage for trained artifacts
  type: gcs
- id: wine
  description: Wine dataset
  type: gcs
```

Quickstart

Congrats! You are now ready to train the model.

Train the model

| | |
|--------|--|
| Before | Project code, hosted on GitHub |
| After | Trained GPPI model (a Trained Model Binary) |

Create payload:

```
$ touch ./odahu-flow/training.odahu.yaml
```

Paste code into the file:

./odahu-flow/training.odahu.yaml

```
1 kind: ModelTraining
2 id: wine-tutorial
3 spec:
4   model:
5     name: wine
6     version: 1.0
7   toolchain: mlflow # MLFlow training toolchain integration
8   entrypoint: main
9   workDir: mlflow/sklearn/wine # MLproject location (in GitHub)
10  data:
11    - connName: wine-tutorial
12      localPath: mlflow/sklearn/wine/wine-quality.csv # wine-quality.csv file
13  hyperParameters:
14    alpha: "1.0"
15  resources:
16    limits:
17      cpu: 4
18      memory: 4Gi
19    requests:
20      cpu: 2
21      memory: 2Gi
22  vcsName: odahu-flow-tutorial
```

In this file, we:

- line 7: Set Odahu toolchain's name to mlflow
- line 8: Reference main method in entry_points (which is defined for MLproject files).
- line 9: Point workDir to the MLFlow project directory. (This is the directory that has the MLproject in it.)
- line 10: A section defining input data
- line 11: connName id of the wine_connection.odahu.yaml (created in the previous step)
- line 12: localPath relative path of the data file at the training (docker) container where data were put
- lines 13-14: Input hyperparameters, defined in MLProject file, and passed to main method

Quickstart

- line 22: vcsName id of the vcs_connection.odahu.yaml (created in the previous step)

Train using Odahu-flow CLI:

```
$ odahuflowctl training create -f ./odahu-flow/training.odahu.yaml
```

Check Train logs:

```
$ odahuflowctl training logs --id wine-tutorial
```

The **Train** process will finish after some time.

To check the status run:

```
$ odahuflowctl training get --id wine-tutorial
```

When the Train process finishes, the command will output this YAML:

- state succeeded
- artifactName (filename of **Trained Model Binary**)

Or Train using the **Plugin for JupyterLab**:

1. Open jupyterlab
2. Open cloned repo, and then the folder with the project
3. Select file ./odahu-flow/training.odahu.yaml and in context menu press submit button

You can see model logs using Odahu cloud mode in the left side tab (cloud icon) in Jupyterlab

1. Open Odahu cloud mode tab
2. Look for TRAINING section
3. Press on the row with *ID=wine*
4. Press button LOGS to connect to **Train** logs

After some time, the **Train** process will finish. Train status is updated in column status of the TRAINING section in the Odahu cloud mode tab. If the model training finishes with success, you will see *status=succeeded*.

Then open **Train** again by pressing the appropriate row. Look at the *Results* section. You should see:

- artifactName (filename of **Trained Model Binary**)

artifactName is the filename of the trained model. This model is in **GPPI** format. We can download it from storage defined in the models-output Connection. (This connection is created during Odahu Platform installation, so we were not required to create this Connection as part of this tutorial.)

Package the model

| | |
|--------|---|
| Before | The trained model in GPPI Trained Model Binary |
| After | Docker image for the packaged model, including a model REST API |

Create payload:

```
$ touch ./odahu-flow/packaging.odahu.yaml
```

Paste code into the file:

```
./odahu-flow/packaging.odahu.yaml
```

```
1 id: wine-tutorial
2 kind: ModelPackaging
3 spec:
4   artifactName: "<fill-in>" # Use artifact name from Train step
5   targets:
6     - connectionName: docker-tutorial # Docker registry where output image will be stored
7       name: docker-push
8   integrationName: docker-rest # REST API Packager
```

In this file, we:

- line 4: Set to artifact name from the Train step
- line 6: Set to docker registry, where output will be staged
- line 7: Specify the docker command
- line 8: id of the **REST API Packager**

Create a **Package** using **Odahu-flow CLI**:

```
$ odahuflowctl packaging create -f ./odahu-flow/packaging.odahu.yaml
```

Check the **Package** logs:

```
$ odahuflowctl packaging logs --id wine-tutorial
```

After some time, the **Package** process will finish.

To check the status, run:

```
$ odahuflowctl packaging get --id wine-tutorial
```

You will see YAML with updated **Package** resource. Look at the status section. You can see:

- **image** # This is the filename of the Docker image in the registry with the trained model prediction, served via REST`.

Or run Package using the **Plugin for JupyterLab**:

Quickstart

1. Open jupyterlab
2. Open the repository that has the source code, and navigate to the folder with the MLProject file
3. Select file ./odahu-flow/packaging.odahu.yaml and in the context menu press the submit button

To view Package logs, use Odahu cloud mode in the side tab of your Jupyterlab

1. Open Odahu cloud mode tab
2. Look for PACKAGING section
3. Click on the row with *ID=wine*
4. Click the button for LOGS and view the Packaging logs

After some time, the **Package** process will finish. The status of training is updated in column status of the *PACKAGING* section in the Odahu cloud mode tab. You should see *status=succeeded*.

Then open PACKAGING again by pressing the appropriate row. Look at the *Results* section. You should see:

- image (this is the filename of docker image in the registry with the trained model as a REST service`);

Deploy the model

| | |
|--------|---|
| Before | Model is packaged as image in the Docker registry |
| After | Model is served via REST API from the Odahu cluster |

Create payload:

```
$ touch ./odahu-flow/deployment.odahu.yaml
```

Paste code into the file:

```
./odahu-flow/deployment.odahu.yaml
```

```
1 id: wine-tutorial
2 kind: ModelDeployment
3 spec:
4   image: "<fill-in>"
5   minReplicas: 1
6   imagePullConnectionID: docker-tutorial
```

In this file, we:

- line 4: Set the image that was created in the Package step
- line 6: Set the id of the **REST API Packager**

Create a **Deploy** using the **Odahu-flow CLI**:

Quickstart

```
$ odahuflowctl deployment create -f ./odahu-flow/deployment.odahu.yaml
```

After some time, the **Deploy** process will finish.

To check its status, run:

```
$ odahuflowctl deployment get --id wine-tutorial
```

Or create a *Deploy* using the **Plugin for JupyterLab**:

1. Open jupyterlab
2. Open the cloned repo, and then the folder with the MLProject file
3. Select file `./odahu-flow/deployment.odahu.yaml`. In context menu press the submit button

You can see Deploy logs using the Odahu cloud mode side tab in your Jupyterlab

1. Open the Odahu cloud mode tab
2. Look for the DEPLOYMENT section
3. Click the row with *ID=wine*

After some time, the **Deploy** process will finish. The status of Deploy is updated in column status of the *DEPLOYMENT* section in the Odahu cloud mode tab. You should see *status=Ready*.

Use the deployed model

| Step input data | The deployed model |
|-----------------|--------------------|
|-----------------|--------------------|

After the model is deployed, you can check its API in Swagger:

Open <your-odahu-platform-host>/swagger/index.html and look and the endpoints:

1. GET `/model/wine-tutorial/api/model/info` - OpenAPI model specification;
2. POST `/model/wine-tutorial/api/model/invoke` - Endpoint to do predictions;

But you can also do predictions using the **Odahu-flow CLI**.

Create a payload file:

```
$ touch ./odahu-flow/r.json
```

Add payload for `/model/wine-tutorial/api/model/invoke` according to the OpenAPI schema. In this payload we provide values for model input variables:

`./odahu-flow/r.json`

```
{  
  "columns": [
```

Quickstart

```
"fixed acidity",
"volatile acidity",
"citric acid",
"residual sugar",
"chlorides",
"free sulfur dioxide",
"total sulfur dioxide",
"density",
"pH",
"sulphates",
"alcohol"
],
"data": [
[
  7,
  0.27,
  0.36,
  20.7,
  0.045,
  45,
  170,
  1.001,
  3,
  0.45,
  8.8
]
]
```

Invoke the model to make a prediction:

```
$ odahuflowctl model invoke --mr wine-tutorial --json-file r.json
```

./odahu-flow/r.json

```
{"prediction": [6.0], "columns": ["quality"]}
```

Congrats! You have completed the tutorial.

API

API service manages Odahu Platform entities.

- **Connections**
- **Trainings**
- **Packaging**
- **Deployments**

API service can provide the following data, when queried:

- Model Train and Deploy logs
- Model **Trainer Metrics**
- Model **Trainer Tags**

API-provided URLs

All information about URLs that **API service** provides can be viewed using the auto-generated, interactive Swagger page. It is located at <api-address>/swagger/index.html. You can read all of the up-to-date documentation and invoke all methods (allowed for your account) right from this web page.

Authentication and authorization

API service analyzes incoming HTTP headers for JWT token, extracts client's scopes from this token and approves / declines incoming requests based on these (provided in JWT) scopes.

Implementation details

API service is a REST server, written in GoLang. For easy integration, it provides a Swagger endpoint with up-to-date protocol information.

| | |
|------------------------------------|--|
| Technologies used | GoLang |
| Distribution representation | Docker Image |
| Source code location | packages/operator |
| Can be used w/o Odahu Platform? | Yes |
| Does it connect to other services? | Yes (Kubernetes API) |
| Can it be deployed locally? | If a local Kubernetes cluster is present |
| Does it provide any interface? | Yes (HTTP REST API) |

Operator

Operator monitors Odahu-provided Kubernetes (K8s) [Custom Resources](#). This gives Operator the ability to manage Odahu entities using K8s infrastructure (Secrets, Pods, Services, etc). The K8s entities that belong to Odahu are referred to as [Odahu-flow's CRDs](#).

Operator is a mandatory component in Odahu clusters.

Implementation details

Operator is a Kubernetes Operator, written using Kubernetes Go packages.

| | |
|--------------------------------------|--|
| Technologies used | GoLang |
| Distribution representation | Docker Image |
| Source code location | packages/operator |
| Can be used w/o Odahu Platform? | Yes |
| Does it connect to another services? | Yes (Kubernetes API) |
| Can be deployed locally? | If local Kubernetes cluster is present |
| Does it provide any interface? | No |

MLFlow Trainer

Odahu provides a **Trainer Extension** for the popular **MLflow** framework.

This allows Python model **Training** in Python, and provides support for MLflow APIs. Trained models are packaged using the **General Python Prediction Interface**.

Limitations

- Odahu supports Python (v. 3) libraries (e.g. Keras, Sklearn, TensorFlow, etc.)
- MLeap is not supported
- Required packages (system and python) must be declared in a conda environment file
- Train must save only one model, using one MLproject entry point method. Otherwise an exception will occur
- Input and output columns should be mapped to the specially-named head_input.pkl and head_output.pkl files to make it into the Packaged artifact
- Training code should avoid direct usage of MLflow client libraries

Implementation Details

| | |
|----------|-------------|
| Support | official |
| Language | Python 3.6+ |

Source code is available on [GitHub](#).

Low-level integration details are provided [here](#).

Metrics

Odahu is pluggable and can integrate with a variety of metrics monitoring tools, allowing monitoring for:

- Model training metrics
- Model performance metrics
- System metrics (e.g. operator counters)

Odahu's installation Helm chart bootstraps a [Prometheus](#) operator to persist metrics and [Grafana](#) dashboard to display them.

Alternative integrations can be similarly constructed that swap in other monitoring solutions.

Airflow

Odahu-flow provides a set of custom operators that allow you to interact with a Odahu cluster using [Apache Airflow](#)

Connections

The Airflow plugin should be authorized by Odahu. Authorization is implemented using regular [Airflow Connections](#)

All custom Odahu-flow operators accept `api_connection_id` as a parameter that refers to *Odahu-flow Connection*

Odahu-flow Connection

The Odahu connection provides access to a Odahu cluster for Odahu custom operators.

Configuring the Connection

Host (required)

The host to connect to. Usually available at: `odahu.<cluster-base-url>`

Type (required)

`HTTP`

Schema (optional)

`https`

Login (required)

Specify the user name to connect.

Password (required)

Specify the password to connect.

Extra (optional)

Specify the extra parameters (as json dictionary) that can be used in Odahu connection. Because Odahu uses OpenID authorization, additional OpenID/OAuth 2.0 parameters may be supplied here.

The following parameters are supported and must be defined:

- **`auth_url`**: url of [authorization server](#)
- **`client_id`**: The client identifier issued to the client during the registration process. [See more](#)
- **`client_secret`**: The client secret. The client MAY omit the parameter if the client secret is an empty string. [See more](#)
- **`scope`**: Access Token Scope

Example “extras” field:

```
{  
    "auth_url": "https://keycloak.<my-app-domain>",  
    "client_id": "my-app",  
    "client_secret": "*****",  
    "scope": "openid profile email offline_access groups",  
}
```

Custom operators

This chapter describes the custom operators provided by Odahu.

Train, Pack, Deploy operators

class TrainingOperator (training=None, api_connection_id=None, *args, **kwargs)

The operator that runs **Train** phase

Use args and kwargs to override other operator parameters

Parameters

- :
 - **training** (*odahuflow.sdk.models.ModelTraining*) - describes the **Train** phase
 - **api_connection_id** (*str*) - conn_id of Odahu-flow Connection

class TrainingSensor (training_id=None, api_connection_id=None, *args, **kwargs)

The operator that waits for **Train** phase is finished

Use args and kwargs to override other operator parameters

Parameters

- :
 - **training_id** (*str*) - Train id waits for
 - **api_connection_id** (*str*) - conn_id of Odahu-flow Connection

class PackagingOperator (packaging=None, api_connection_id=None, trained_task_id: str = "", *args, **kwargs)

The operator that runs **Package** phase

Use args and kwargs to override other operator parameters

Parameters

- :
 - **packaging** (*odahuflow.sdk.models.ModelPackaging*) - describes the **Package** phase
 - **api_connection_id** (*str*) - conn_id of Odahu-flow Connection
 - **trained_task_id** (*str*) - finished task id of TrainingSensor

class PackagingSensor (training_id=None, api_connection_id=None, *args, **kwargs)

The operator that waits for **Package** phase is finished

Use args and kwargs to override other operator parameters

Custom operators

- Parameters :
 - **packaging_id** (str) - Package id waits for
 - **api_connection_id** (str) - conn_id of Odahu-flow Connection

`class DeploymentOperator` (deployment=None, api_connection_id=None, *args, **kwargs)

The operator that runs **Deploy** phase

Use args and kwargs to override other operator parameters

- Parameters :
 - **packaging** (`odahuflow.sdk.models.ModelDeployment`) - describes the **Deploy** phase
 - **api_connection_id** (str) - conn_id of Odahu-flow Connection
 - **packaging_task_id** (str) - finished task id of PackagingSensor

`class DeploymentSensor` (training_id=None, api_connection_id=None, *args, **kwargs)

The operator that waits for **Deploy** phase is finished

Use args and kwargs to override other operator parameters

- Parameters :
 - **deployment_id** (str) - Deploy id waits for
 - **api_connection_id** (str) - conn_id of Odahu-flow Connection

Model usage operators

These operators are used to interact with deployed models.

`class ModelInfoRequestOperator` (self, model_deployment_name: str, api_connection_id: str, model_connection_id: str, md_role_name: str = "", *args, **kwargs)

The operator what extract metadata of deployed model.

Use args and kwargs to override other operator parameters

- Parameters :
 - **model_deployment_name** (str) - Model deployment name
 - **api_connection_id** (str) - conn_id of Odahu-flow Connection
 - **model_connection_id** (str) - id of Odahu **Connection** for deployed model access
 - **md_role_name** (str) - Role name

`class ModelPredictRequestOperator` (self, model_deployment_name: str, api_connection_id: str, model_connection_id: str, request_body: typing.Any, md_role_name: str = "", *args, **kwargs)

The operator request prediction using deployed model.

Use args and kwargs to override other operator parameters

How to describe operators

Parameters :

- **model_deployment_name** (str) - <paste>
- **api_connection_id** (str) - conn_id of Odahu-flow Connection
- **model_connection_id** (str) - id of Odahu **Connection** for deployed model access
- **request_body** (dict) - JSON Body with model parameters
- **md_role_name** (str) - Role name

Helper operators

These operators are helpers to simplify using Odahu-flow.

```
class GcpConnectionToOdahuConnectionOperator(self, api_connection_id: str,
google_cloud_storage_conn_id: str, conn_template: typing.Any, *args, **kwargs)
    Create Odahu-flow Connection using GCP Airflow Connection
    Use args and kwargs to override other operator parameters
```

Parameters :

- **api_connection_id** (str) - conn_id of Odahu-flow Connection
- **google_cloud_storage_conn_id** (str) - conn_id to Gcp Connection
- **conn_template**
(*odahuflow.sdk.models.connection.Connection*) - Odahu-flow Connection template

How to describe operators

When you initialize Odahu custom operators such as TrainingOperator, PackagingOperator, or DeploymentOperator you should pass odahu resource payload as a parameter.

Actually, this is payload that describes a resource that will be created at Odahu-flow cluster. You should describe such payloads using *odahuflow.sdk* models

Creating training payload

```
training = ModelTraining(
    id=training_id,
    spec=ModelTrainingSpec(
        model=ModelIdentity(
            name="wine",
            version="1.0"
        ),
        toolchain="mlflow",
        entrypoint="main",
        work_dir="mlflow/sklearn/wine",
        hyper_parameters={
            "alpha": "1.0"
        },
        data=[
            DataBindingDir(
                path="path/to/dataset"
            )
        ]
    )
)
```

How to describe operators

```
        conn_name='wine',
        local_path='mlflow/sklearn/wine/wine-quality.csv'
    ),
],
resources=ResourceRequirements(
    requests=ResourceList(
        cpu="2024m",
        memory="2024Mi"
    ),
    limits=ResourceList(
        cpu="2024m",
        memory="2024Mi"
    )
),
vcs_name="odahu-flow-examples"
),
)
```

But if you did some RnD work with Odahu-flow previously, it's likely that you already have yaml/json files that describe the same payloads. You can reuse them to create odahuflow.sdk models automatically

Using plain yaml/json text

```
from odahuflow.airflow.resources import resource

packaging_id, packaging = resource("""
id: airflow-wine
kind: ModelPackaging
spec:
    artifactName: "<fill-in>"
    targets:
        - connectionName: docker-ci
          name: docker-push
          integrationName: docker-rest
"""
)"""
```

Or refer to yaml/json files that must be located at Airflow DAGs folder or Airflow Home folder (these folders are configured at airflow.cfg file)

Creating training payload

```
from odahuflow.airflow.resources import resource
training_id, training = resource('training.odahuflow.yaml')
```

In this file, we refer to file *training.odahuflow.yaml* that is located at airflow dag's folder

For example, if you use [Google Cloud Composer](#) then you can locate your yaml's inside DAGs bucket and refer to them by relative path:

```
gsutil cp ~/.training.odahuflow.yaml gs://<your-composer-dags-bucket>/
```

DAG example

The example of the DAG that uses custom Odahu-flow operators is shown below. Four DAGs are described.

How to describe operators

dag.py

```
1 from datetime import datetime
2 from airflow import DAG
3 from airflow.contrib.operators.gcs_to_gcs import GoogleCloudStorageToGoogleCloudStorageOperator
4 from airflow.models import Variable
5 from airflow.operators.bash_operator import BashOperator
6 from odahuflow.sdk.models import ModelTraining, ModelTrainingSpec, ModelIdentity, ResourceRequirements, ResourceList, \
7     ModelPackaging, ModelPackagingSpec, Target, ModelDeployment, ModelDeploymentSpec, Connection, ConnectionSpec, \
8     DataBindingDir
9
10 from odahuflow.airflow.connection import GcpConnectionToOdahuConnectionOperator
11 from odahuflow.airflow.deployment import DeploymentOperator, DeploymentSensor
12 from odahuflow.airflow.model import ModelPredictRequestOperator, ModelInfoRequestOperator
13 from odahuflow.airflow.packaging import PackagingOperator, PackagingSensor
14 from odahuflow.airflow.training import TrainingOperator, TrainingSensor
15
16 default_args = {
17     'owner': 'airflow',
18     'depends_on_past': False,
19     'start_date': datetime(2019, 9, 3),
20     'email_on_failure': False,
21     'email_on_retry': False,
22     'end_date': datetime(2099, 12, 31)
23 }
24
25 api_connection_id = "odahuflow_api"
26 model_connection_id = "odahuflow_model"
27
28 gcp_project = Variable.get("GCP_PROJECT")
29 wine_bucket = Variable.get("WINE_BUCKET")
30
31 wine_conn_id = "wine"
32 wine = Connection(
33     id=wine_conn_id,
34     spec=ConnectionSpec(
35         type="gcs",
36         uri=f'gs://{wine_bucket}/data/wine-quality.csv',
37         region=gcp_project,
38     )
39 )
40
41 training_id = "airflow-wine"
42 training = ModelTraining(
43     id=training_id,
44     spec=ModelTrainingSpec(
45         model=ModelIdentity(
46             name="wine",
47             version="1.0"
48         ),
49         toolchain="mlflow",
50         entrypoint="main",
51         work_dir="mlflow/sklearn/wine",
52         hyper_parameters={
53             "alpha": "1.0"
54         },
55         data=[
56             DataBindingDir(
57                 conn_name='wine',
58                 local_path='mlflow/sklearn/wine/wine-quality.csv'
59             ),
60         ],
61         resources=ResourceRequirements(
62             requests=ResourceList(
63                 cpu="2024m",
64                 memory="2024Mi"
65             ),
66             limits=ResourceList(
67                 cpu="2024m",
68                 memory="2024Mi"
69             )
70         ),
71         vcs_name="odahu-flow-examples"
72     ),
73 )
74
75 packaging_id = "airflow-wine"
76 packaging = ModelPackaging(
77     id=packaging_id,
78     spec=ModelPackagingSpec(
79         targets=[Target(name="docker-push", connection_name="docker-ci")],
80         integration_name="docker-rest"
```

```
72     ),
73 )
74
75 packaging_id = "airflow-wine"
76 packaging = ModelPackaging(
77     id=packaging_id,
78     spec=ModelPackagingSpec(
79         targets=[Target(name="docker-push", connection_name="docker-ci")],
80         integration_name="docker-rest"
```

How to describe operators

```
81     ),
82 )
83
84 deployment_id = "airflow-wine"
85 deployment = ModelDeployment(
86     id=deployment_id,
87     spec=ModelDeploymentSpec(
88         min_replicas=1,
89     ),
90 )
91
92 model_example_request = {
93     "columns": ["alcohol", "chlorides", "citric acid", "density", "fixed acidity", "free sulfur dioxide", "pH",
94                 "residual sugar", "sulphates", "total sulfur dioxide", "volatile acidity"],
95     "data": [[12.8, 0.029, 0.48, 0.98, 6.2, 29, 3.33, 1.2, 0.39, 75, 0.66],
96              [12.8, 0.029, 0.48, 0.98, 6.2, 29, 3.33, 1.2, 0.39, 75, 0.66]]
97 }
98
99 dag = DAG(
100     'wine_model',
101     default_args=default_args,
102     schedule_interval=None
103 )
104
105 with dag:
106     data_extraction = GoogleCloudStorageToGoogleCloudStorageOperator(
107         task_id='data_extraction',
108         google_cloud_storage_conn_id='wine_input',
109         source_bucket=wine_bucket,
110         destination_bucket=wine_bucket,
111         source_object='input/*.csv',
112         destination_object='data/',
113         project_id=gcp_project,
114         default_args=default_args
115     )
116     data_transformation = BashOperator(
117         task_id='data_transformation',
118         bash_command='echo "imagine that we transform a data"',
119         default_args=default_args
120     )
121     odahuflow_conn = GcpConnectionToOdahuConnectionOperator(
122         task_id='odahuflow_connection_creation',
123         google_cloud_storage_conn_id='wine_input',
124         api_connection_id=api_connection_id,
125         conn_template=wine,
126         default_args=default_args
127     )
128
129     train = TrainingOperator(
130         task_id="training",
131         api_connection_id=api_connection_id,
132         training=training,
133         default_args=default_args
134     )
135
136     wait_for_train = TrainingSensor(
137         task_id='wait_for_training',
138         training_id=training_id,
139         api_connection_id=api_connection_id,
140         default_args=default_args
141     )
142
143     pack = PackagingOperator(
144         task_id="packaging",
145         api_connection_id=api_connection_id,
146         packaging=packaging,
147         trained_task_id="wait_for_training",
148         default_args=default_args
149     )
150
151     wait_for_pack = PackagingSensor(
152         task_id='wait_for_packaging',
153         packaging_id=packaging_id,
```

```
154         api_connection_id=api_connection_id,
155         default_args=default_args
156     )
157
158     dep = DeploymentOperator(
159         task_id="deployment",
160         api_connection_id=api_connection_id,
161         deployment=deployment,
162         packaging_task_id="wait_for_packaging",
```

How to describe operators

```
163     default_args=default_args
164 )
165
166 wait_for_dep = DeploymentSensor(
167     task_id='wait_for_deployment',
168     deployment_id=deployment_id,
169     api_connection_id=api_connection_id,
170     default_args=default_args
171 )
172
173 model_predict_request = ModelPredictRequestOperator(
174     task_id="model_predict_request",
175     model_deployment_name=deployment_id,
176     api_connection_id=api_connection_id,
177     model_connection_id=model_connection_id,
178     request_body=model_example_request,
179     default_args=default_args
180 )
181
182 model_info_request = ModelInfoRequestOperator(
183     task_id='model_info_request',
184     model_deployment_name=deployment_id,
185     api_connection_id=api_connection_id,
186     model_connection_id=model_connection_id,
187     default_args=default_args
188 )
189
190 data_extraction >> data_transformation >> odahuflow_conn >> train
191 train >> wait_for_train >> pack >> wait_for_pack >> dep >> wait_for_dep
192 wait_for_dep >> model_info_request
193 wait_for_dep >> model_predict_request
```

In this file, we create four dags:

- DAG on line 190 extract and transform data, create Odahu-flow connection and run **Train**
- DAG on line 191 sequentially run phases **Train, Package, Deploy**
- DAG on line 192 wait for model deploy and then extract schema of model predict API
- DAG on line 193 wait for model deploy and then invoke model prediction API

JupyterLab extension

Odahu-flow provides the JupyterLab extension that allows you to interact with an Odahu cluster from JupyterLab web-based IDEs.

Installation

To install the extension, perform the following steps:

```
pip install odahu-flow-jupyterlab-plugin
jupyter serverextension enable --sys-prefix --py odahuflow.jupyterlab
jupyter labextension install odahu-flow-jupyterlab-plugin
```

Another option is [prebuilt Jupyterlab Docker Image](#) with the extension.

Configuration

The extension can be configured through the environment variables.

| Environment name | Default | Value example | Description |
|------------------------------|---------|---|---|
| DEFAULT_API_ENDPOINT | | https://odahu.company.com/ | Default URL to the Odahu-flow API server |
| API_AUTH_ENABLED | true | true | Change the value to false if authorization is disabled on the Odahu-flow API server |
| ODAHUFLOWCTL_OAUTH_AUTH_URL | | https://keycloak.company.org/auth/realms/master/protocol/openid-connect/auth | Keycloak authorization endpoint |
| JUPYTER_REDIRECT_URL | | http://localhost:8888 | JupyterLab external URL |
| ODAHUFLOWCTL_OAUTH_CLIENT_ID | | | Oauth client ID |

Login

| | | | |
|--|--|--|----------------------|
| ODAHUF LOWCTL_ OAUTH_ CLIENT_S ECRET | | | Oauth2 client secret |
|--|--|--|----------------------|

To enable SSO, you should provide the following options:

- *ODAHUFLOWCTL_OAUTH_AUTH_URL*
- *JUPYTER_REDIRECT_URL*
- *ODAHUFLOWCTL_OAUTH_CLIENT_SECRET*
- *ODAHUFLOWCTL_OAUTH_CLIENT_ID*

Login

To authorize on an Odahu-flow API service in the Jupyterlab extension, you should perform the following steps:

- Copy and paste the Odahu-flow API service URL.
- Open an API server URL in a browser to get the token. Copy and paste this token in the login form.

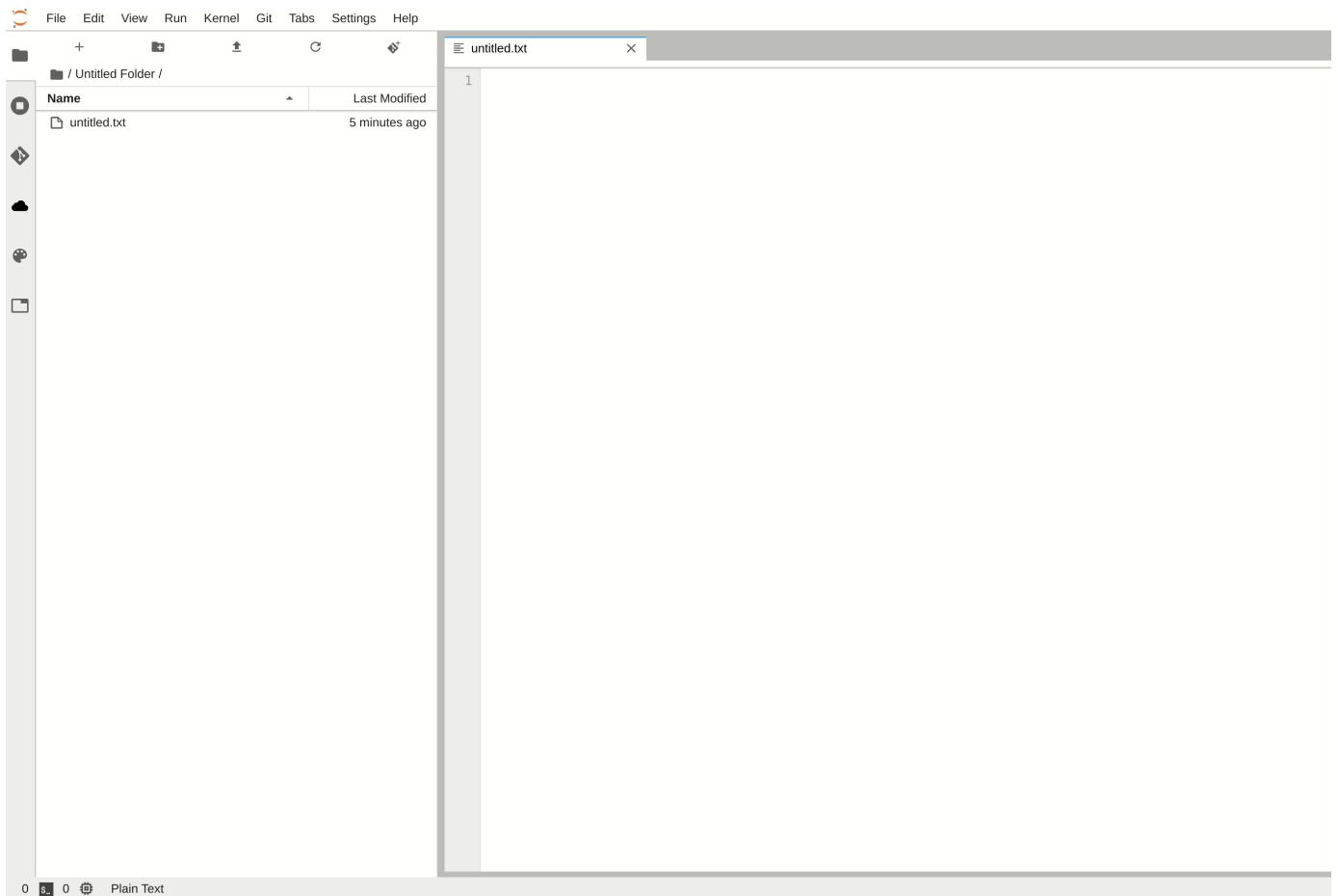
Usage

Below we consider all views of the JupyterLab extension.

Templates

The extension provides predefined list of API file templates. You can create a file from a template.

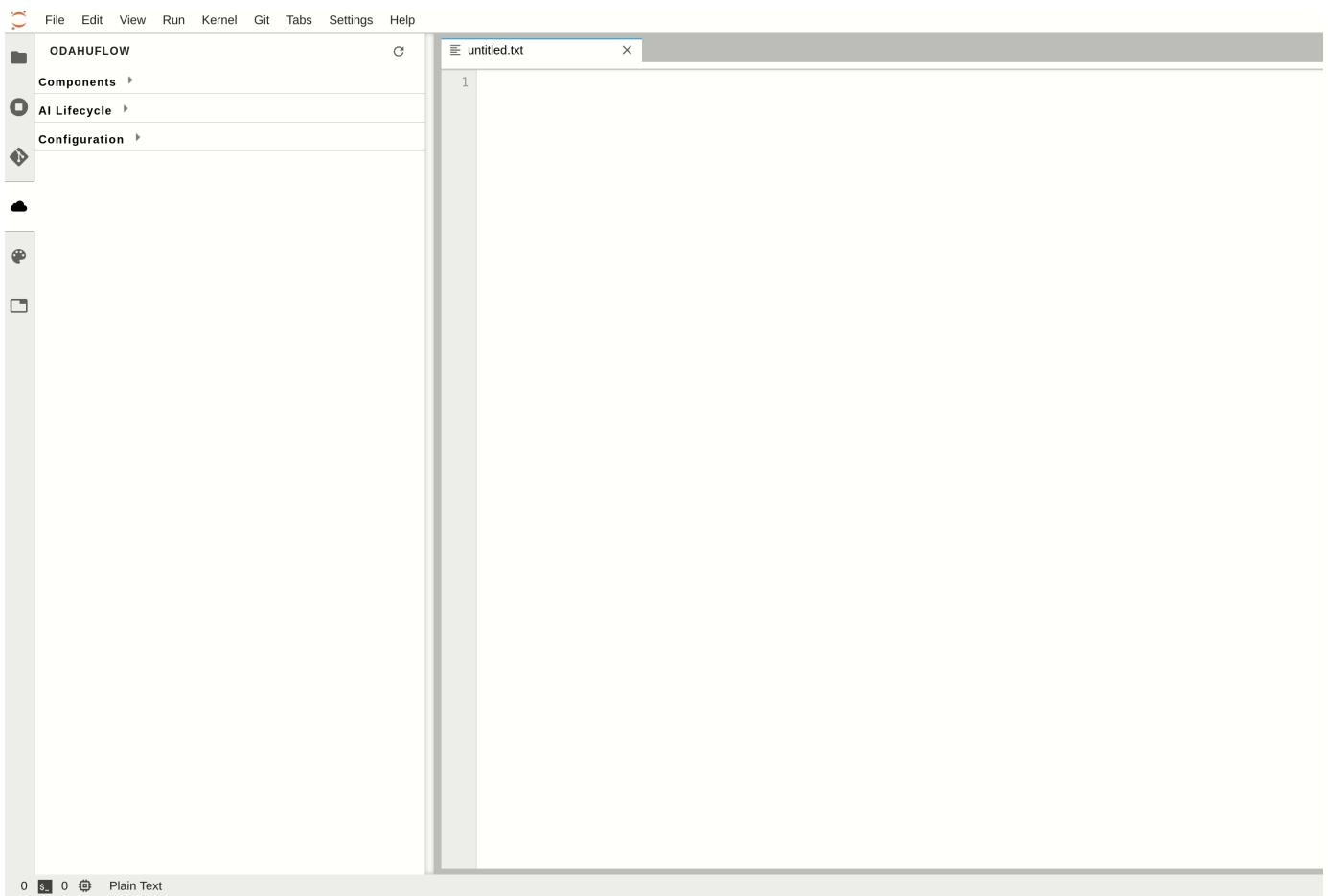
Login



Main view

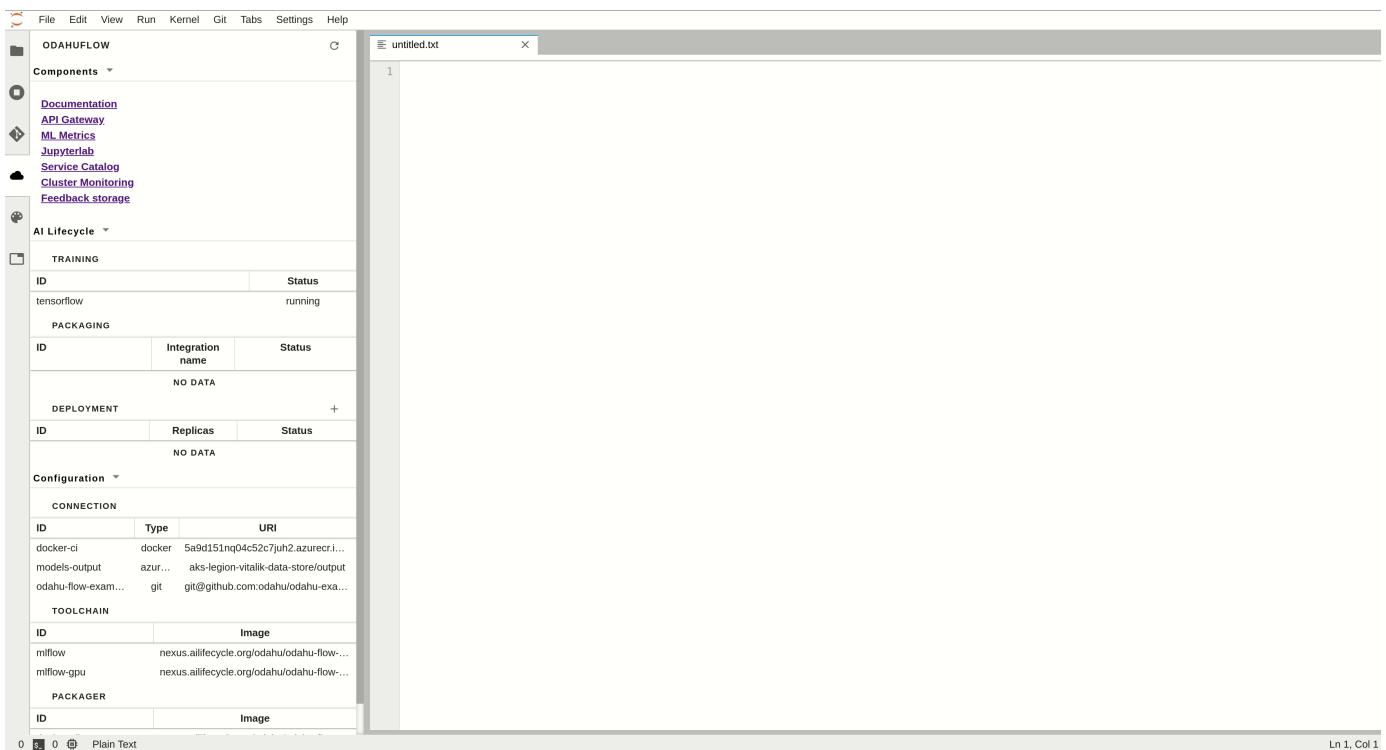
The main view contains all Odahu-flow entities. You can view or delete them.

Login



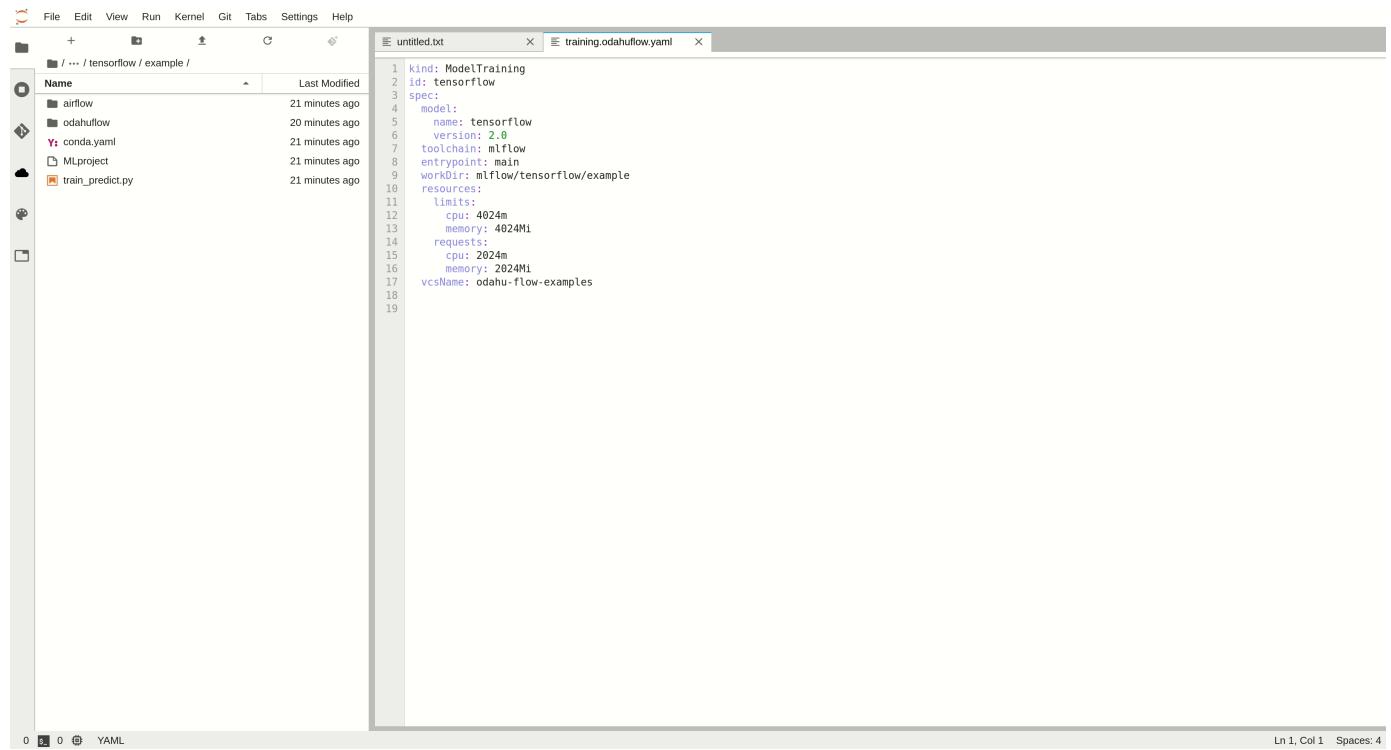
Log viewer

For troubleshooting, you can get access to the training, packaging or deployment logs. If the job is running then logs will be updated in runtime.



Submit resources

You can create any Odahu-flow entities from the extension. The button **Submit** only appears in the context menu when file ends with .yaml or json.



The screenshot shows the Odahu Flow extension interface. On the left is a file explorer sidebar with a tree view of files in the directory `/ ... / tensorflow / example /`. The files listed are `airflow`, `odahuflow`, `conda.yaml` (highlighted in yellow), `MLproject`, and `train_predict.py`. The main area contains two code editors. The left editor has the tab `untitled.txt` and displays the following YAML configuration:

```
1 kind: ModelTraining
2 id: tensorflow
3 spec:
4   model:
5     name: tensorflow
6     version: 2.0
7     toolchain: mlflow
8     entrypoint: main
9     workDir: mlflow/tensorflow/example
10    resources:
11      limits:
12        cpu: 4024m
13        memory: 4024Mi
14      requests:
15        cpu: 2024m
16        memory: 2024Mi
17    vcsName: odahu-flow-examples
18
19
```

The right editor has the tab `training.odahuflow.yaml` and is currently empty. At the bottom of the interface, there are status indicators: 0 files, 0 errors, and 1 warning, followed by the text "YAML". On the far right, there is a status bar with the text "Ln 1, Col 1 Spaces: 4".

Odahuflowctl

Odahuflowctl (`odahuflowctl`) is a command-line interface for interacting with Odahu-flow API service.

Installation

Odahu-flow CLI is available in PyPi repository. You should execute the following command to install `odahuflowctl`:

```
pip install odahu-flow-cli  
odahuflowctl --version
```

Help

To read `odahuflowctl` help, you should use the following command:

```
odahuflowctl --help
```

for a specific command, for example, get list of model deployments:

```
odahuflowctl deployment get --help
```

Login

There are two authentication types for Odahu CLI.

Specifying of a token explicitly

You should open an API server URL in a browser to get the login command. The command already contains your token. Copy and paste provided command into your shell.

Example of command:

```
odahuflowctl login --url <api-url> --token <your-token>
```

Sign in interactively

This method will use a web browser to sign in.

Odahuflowctl

Run the login command:

```
odahuflowctl login --url <api-url>
```

Odahu CLI will open an IAM server in your default browser. Sign in with your account credentials.

Model Format

The **Odahu Model Artifact Format** (OMAF) describes a format to package, store, and transport ML models.

Models can be built in different languages and use different platform libraries. For example: {Python, Scala, R, ...} using {scikit-learn, tensorflow, keras, ...}.

An OMAF **Artifact** is stored as a file-system folder packed into a ZIP file using the Deflate ZIP compression algorithm.

The Artifact contains:

- `odahuflow.model.yaml` a YAML file in the root folder. This file contains meta-information about the type of binary model and other model related information (e.g. language, import endpoints, dependencies).
- Additional folders and files, depending upon meta-information declared in `odahuflow.model.yaml`.

odahuflow.model.yaml

File structure:

- `binaries` - Language and dependencies that should be used to load model binaries
- `binaries.type` - Required Odahu Model Environments. See section [Odahu Model Environments](#).
- `binaries.dependencies` - Dependency management system, compatible with the selected Odahu Model Environment
- `binaries.<additional>` - Model Environment and dependency management system values, for example 'a path to the requirements file'
- `model` - Location of the model artifact Model artifact format depends on [Odahu Model Environment](#).
- `model.name` - name of the model, [a-Z0-9-]+
- `model.version` - version of model. Format is <Apache Version>-<Additional suffix>, where Additional suffix is a [a-Z0-9-.]+ string.
- `model.workDir` - working directory to start model from.
- `model.entrypoint` - name of model artifact (e.g. Python module or Java JAR file).
- `odahuflowVersion` - OMAF version
- `toolchain` - toolchain used for training and preparing the Artifact
- `toolchain.name` - name of the toolchain
- `toolchain.version` - version of used toolchain.
- `toolchain.<additional>` - additional fields, related to used toolchain (e.g. used submodule of toolchain).

Odahu Model Environments

Examples:

Example with GPPI using conda for dependency management, mlflow toolchain.

```
binaries:
  type: python
  dependencies: conda
  conda_path: mlflow/model/mlflow_env.yml
model:
  name: wine-quality
  version: 1.0.0-12333122
  workDir: mlflow/model
  entrypoint: entrypoint
  odahuflowVersion: '1.0'
toolchain:
  name: mlflow
  version: 1.0.0
```

Odahu Model Environments

Odahu supports these model environments:

- General Python Prediction Interface (GPPI). Can import a trained model as a python module and use a predefined function for prediction. Value for binaries.type should be python.
- General Java Prediction Interface (GJPI). Can import a trained model as a Java Library and use a predefined interfaces for prediction. Value for binaries.type should be java.

Odahu's General Python Prediction Interface (GPPI)

General Information

Description

This interface is an importable Python module with a declared interface (functions with arguments and return types). Toolchains that save models in this format must provide an entrypoint with this interface or they may provide a wrapper around their interface for this interface.

Required Environment variables

- MODEL_LOCATION – path to model's file, relative to working directory.

Interface declaration

Interface functions:

Connections

Odahu needs to know how to connect to a bucket, git repository, and so on. This kind of information is handled by Connection API.

General connection structure

All types of connections have the same general structure. But different connections require a different set of fields. You can find the examples of specific type of connection in the id of the Connection types section. Below you can find the description of all fields:

Connection API

```
# Unique value among all connections
id: "id-12345"
spec:
  # Optionally description of a connection
  description: "Some description"
  # Optionally link to the web resource. For example, git repo or a gcp bucket
  webUILink: https://test.org/123
  # URI. It is a required value.
  uri: s3://some-bucket/path/file
  # Type of a connection. Available values: s3, gcs, azureblob, git, docker, ecr.
  type: s3
  # Username
  username: admin
  # Password
  password: admin
  # Service account role
  role: some-role
  # AWS region or GCP project
  region: some region
  # VCS reference
  reference: develop
  # Key ID
  keyID: "1234567890"
  # SSH or service account secret
  keySecret: b2RhaHUK
  # SSH public key
  publicKey: b2RhaHUK
```

Connection management

Connections can be managed using the following ways.

Swagger UI

Swagger UI is available at <http://api-service/swagger/index.html> URL.

Odahu-flow CLI

Connection types

Odahuflowctl supports the connection API. You must be login if you want to get access to the API.

- Getting all connections in json format:

```
odahuflowctl conn get --format json
```

- Getting the reference of the connection:

```
odahuflowctl conn get --id odahu-flow-examples -o 'jsonpath=[*].spec.reference'
```

- Creating of a connection from *conn.yaml* file:

```
odahuflowctl conn create -f conn.yaml
```

- All connection commands and documentation:

```
odahuflowctl conn --help
```

JupyterLab

Odahu-flow provides the JupyterLab extension for interacting with Connection API.

Connection types

For now, Odahu-flow supports the following connections types:

- S3
- Google Cloud Storage
- Azure Blob storage
- GIT
- Docker
- Amazon Elastic Container Registry

S3

An S3 connection allows interactions with s3 API. This type of connection is used as storage of:

- model trained artifacts.
- input data for ML models.

Connection types

Note

You can use any S3 compatible API, for example minio or Ceph.

Before usage, make sure that:

- You have created an AWS S3 bucket. [Examples of Creating a Bucket](#).
- You have created an IAM user that has access to the AWS S3 bucket. [Creating an IAM User in Your AWS Account](#).
- You have created the IAM keys for the user. [Managing Access Keys for IAM Users](#).

Note

At that moment, Odahu-flow only supports authorization through [IAM User](#). We will support AWS service role and authorization using temporary credentials in the near future.

The following fields of connection API are required:

- spec.type - It must be equal **s3**.
- spec.keyID - an access key ID (for example, AKIAIOSFODNN7EXAMPLE).
- spec.keySecret - a secret access key (for example, wJalrXUtnFEMI/K7MDENG/bPxRfiCYEXAMPLEKEY).
- spec.uri - S3 compatible URI, for example s3://<bucket-name>/dir1/dir2/
- spec.region - [AWS Region](#), where a bucket was created.

Example of Connection S3:

```
id: "training-data"
spec:
  type: s3
  uri: s3://raw-data/model/input
  keyID: "AKIAIOSFODNN7EXAMPLE"
  keySecret: "wJalrXUtnFEMI/K7MDENG/bPxRfiCYEXAMPLEKEY"
  description: "Training data for a model"
  region: eu-central-1
```

Google Cloud Storage

[Google Cloud Storage](#) allows storing and accessing data on Google Cloud Platform infrastructure. This type of connection is used as storage of:

- model trained artifacts.
- input data for ML models.

Before usage, make sure that:

Connection types

- You have created an GCS bucket. [Creating storage buckets](#).
 - You have created an service account. [Creating and managing service accounts](#).
 - You have assigned roles/storage.objectAdmin role on the service account for the GCS bucket. [Using Cloud IAM permissions](#).
 - You have created the IAM keys for the service account. [Creating and managing service account keys](#).

Note

Workload Identity is the recommended way to access Google Cloud services from within GKE due to its improved security properties and manageability. We will support the Workload Identity in the near future.

The following fields of connection API are required:

- `spec.type` - It must be equal **gcs**.
 - `spec.keySecret` - a service account key in json format.
 - `spec.uri` - GCS compatible URI, for example `gcs://<bucket-name>/dir1/dir2/`
 - `spec.region` - **GCP Region**, where a bucket was created.

Azure Blob storage

Odahu-flow uses the [Blob storage](#) in Azure to store:

- model trained artifacts.
 - input data for ML models.

Before usage, make sure that:

- You have created a storage account . [Create a storage account.](#)
 - You have created a storage container in the storage account . [Create a container.](#)
 - You have created a SAS token. [Create an account SAS.](#)

The following fields of connection API are required:

- `spec.type` - It must be equal **azureblob**.
 - `spec.keySecret` - Odahu-flow uses the **shared access signatures** to authorize in Azure. The key has the following format:
“`<primary_blob_endpoint>/<sas_token>`”.
 - `spec.uri` - Azure storage compatible URI, for example
`<bucket-name>/dir1/dir2/`

Connection types

Example of Connection Blob Storage:

```
id: "training-data"
spec:
  type: azureblob
  container: model/input
  keySecret: https://myaccount.blob.core.windows.net/?restype=service&comp=properties&sv=2019-02-02&ss=bf&sr=s&st=2019-08-01T22%3A18%3A26Z&se=2019-08-10T02%3A23%3A26Z&sr=b&sp=rw&sig=168.1.5.60-168.1.5.70&spr=https&sig=F%6GRVAZ5Cdj2Pw4tgI711STk%gn7bUkkAgB%6HESXmf%4B
```

GIT

Odahu-flow uses the GIT type connection to download a ML source code from a git repository.

The following fields of connection API are required:

- spec.type - It must be equal **git**.
- spec.keySecret - a base64 encoded SSH private key.
- spec.uri - GIT SSH URL, for example
<git@github.com:odahu/odahu-examples.git>
- spec.reference - a branch, tag, or commit.

Example of command to encode ssh key:

```
cat ~/.ssh/id_rsa | base64 -w0
```

Note

Odahu-flow only supports authorization through SSH.

Warning

We recommend using the read-only deploy keys: [Github docs](#) or [Gitlab docs](#).

Example of GIT Connection:

```
id: "ml-repository"
spec:
  type: git
  uri: git@github.com:odahu/odahu-examples.git
  keySecret: C1NVUEVSIFNFQ1JFVAoK
  reference: master
  description: "Git repository with the Odahu-Flow examples"
  webUILink: https://github.com/odahu/odahu-examples
```

Docker

This type of connection is used for pulling and pushing of the Odahu packager result Docker images to a Docker registry. We have been testing the following Docker repositories:

Connection types

- Docker Hub
- Nexus
- Google Container Registry
- Azure Container Registry

Warning

Every docker registry has its authorization specificity. But you must be able to authorize by a username and password. Read the documentation.

Before usage, make sure that:

- You have a username and password.

The following fields of connection API are required:

- `spec.type` - It must be equal **docker**.
- `spec.username` - docker registry username.
- `spec.password` - docker registry password.
- `spec.uri` - docker registry host.

Warning

Connection URI must not contain an URI schema.

Example of GCR Docker connection

```
id: "gcr-registry"
spec:
  type: docker
  uri: https://gcr.io/odahu/
  username: "username"
  password: "password"
```

Example of Docker Hub

```
id: "docker-registry"
spec:
  type: docker
  uri: docker.io/odahu/
  username: "username"
  password: "password"
```

Amazon Elastic Container Registry

Amazon Elastic Container Registry is a managed AWS Docker registry. This type of connection is used for pulling and pushing of the Odahu packager result Docker images.

Note

The Amazon Docker registry does not support a long-lived credential and requires explicitly to create a repository for every image. These are the reasons why we create a dedicated type of connection for the ECR.

Before usage, make sure that:

- You have created an ECR repository. [Creating an ECR Repository](#).
- You have created an IAM user that has access to the ECR repository. [Creating an IAM User in Your AWS Account](#).
- You have created the IAM keys for the user. [Managing Access Keys for IAM Users](#).

The following fields of connection API are required:

- `spec.type` - It must be equal **ecr**.
- `spec.keyID` - an access key ID (for example, AKIAIOSFODNN7EXAMPLE).
- `spec.keySecret` - a secret access key (for example, wJalrXUtnFEMI/K7MDENG/bPxRfCYEXAMPLEKEY).
- `spec.uri` - The url must have the following format,
`aws_account_id.dkr.ecr.`region`.amazonaws.com/some-prefix`.
- `spec.region` - [AWS Region](#), where a docker registry was created.

Example of Connection ECR:

```
id: "docker-registry"
spec:
  type: ecr
  uri: 5555555555.dkr.ecr.eu-central-1.amazonaws.com/odahuflow
  keyID: "AKIAIOSFODNN7EXAMPLE"
  keySecret: "wJalrXUtnFEMI/K7MDENG/bPxRfCYEXAMPLEKEY"
  description: "Packager registry"
  region: eu-central-1
```

Model Trainings

The Odahu-flow Model Training API provides features to manage remote training jobs. A primary goal of API creates a **Trained Model Binary** for a **Packager**. The API is pluggable and can be extended for different ML frameworks.

You can find the list of out-of-the-box trainers below:

- MLFlow

General training structure

Training API

```

kind: ModelTraining
# Some unique value among all trainings
id: wine-12345
spec:
  model:
    # Human-readable model name
    name: wine
    # Human-readable model version
    version: 3.0
    # Optionally, you can specify template for output artifact
    # The default value is {{ .Name }}-{{ .Version }}-{{ .RandomUUID }}.zip
    # where:
    #   Name - spec.model.name
    #   Version - spec.model.version
    #   RandomUUID - a random UUID v4, for example be17d12d-df43-4588-99e7-56a0db3cad77
    artifactNameTemplate: {{ .Name }}-{{ .Version }}-{{ .RandomUUID }}.zip
# The toolchain parameter is a point of extension for different ML frameworks.
# For now, we only support the Mlflow toolchain
toolchain: mlflow
# Mlflow MLProject file can contains the list of entrypoints. You must choose one of these.
entrypoint: main
# Working directory inside a GIT repository
workDir: work/dir
# The training data for a ML script. You can find find full description :ref:`here <ref_trainings:Training Data>`.
data:
  # You can specify a connection name
  - connName: wine
    # Local path with file or dir
    localPath: mlflow/wine-quality/
    # Path to the dir or file in a bucket
    # Optional. If it is missing then the path from connection will be used.
    remotePath: training-data/
# You can specify the map of hyperparameters
hyperParameters:
  key: value
  var2: test
# Compute resources for the training job.
resources:
  limits:
    cpu: 1
    memory: 1Gi
  requests:
    cpu: 1
    memory: 1Gi
# Custom environment variables that should be set before entrypoint invocation.
envs:
  # The name of variable
  - name: TEST_ENV_KEY
    # The value of variable
    value: TEST_ENV_VALUE
# A Docker image where the training will be launched.

```

```

# By default, the image from a toolchain is used.
# image: python:3.8
# A connection which describes credentials to a GIT repository

```

Training data

```
vcsName: <git-connection>
status:
# One of the following states: scheduling, running, succeeded, failed, unknown
state: running
# List of training results
artifacts:
# Mlflow run ID
- runId: 12345678
# Trained artifact name
artifactName: wine-10.zip
# VCS commit ID
commitID: d3d6e8ed776ed37fd2efd7a1b8d5fabdd7e3eea5
```

Training data

Odahu-flow allows downloading data from various sources to the local file system of a training job. Data source supports the following types of Odahu-flow connections:

- S3
 - Google Cloud Storage
 - Azure Blob storage

Let's consider the following example of downloading training data from Google Cloud Storage.

Prerequisites:

- The training data set is located in the *wine-training-data* bucket by *wine/11-11-2011/* directory.
 - The ML script expects that the data will be located in *data/* directory relative to the root git directory.

First of all, we should create an Odahu-flow GCS connection.

Finally, we provide a data section of Model Training.

Example of Connection GCS:

```
spec:  
  data:  
    - connName: wine-training-data-conn  
      localPath: data/  
      remotePath: wine/11-11-2011/
```

Trainings management

Trainings can be managed using the following ways.

Swagger UI

ModelTraining and ToolchainIntegration are available on the Swagger UI at <http://api-service/swagger/index.html> URL.

Odahu-flow CLI

Odahuflowctl supports the Training API. You must be logged in if you want to get access to the API.

Getting all trainings in json format:

```
odahuflowctl train get --format json
```

Getting the model name of the trainings:

```
odahuflowctl train get --id tensorflow-cli -o 'jsonpath=[*].spec.model.name'
```

- Creating a training from *train.yaml* file:

```
odahuflowctl train create -f train.yaml
```

- All training commands and documentation:

```
odahuflowctl train --help
```

JupyterLab

Odahu-flow provides the JupyterLab extension for interacting with Training API.

MLFlow

MLflow is library-agnostic. You can use it with any machine learning library, and in any programming language, since all functions are accessible through a REST API and CLI.

Installation

The most straightforward way to install the MLFlow trainer on an Odahu Cluster is to deploy the *odahu-flow-mlflow* helm chart. The helm chart registers the trainer in the API Odahu and deploys an a **MLflow Tracking Server**. By default, the deployed MLflow Tracking Server is available at <https://cluster-url/mlflow> address.

```
# Add the odahu-flow helm repository
helm repo add odahu-flow 'https://raw.githubusercontent.com/odahu/odahu-helm/master/'
helm repo update
# Fill in the values for the chart or leave the default values
helm inspect values odahu-flow/odahu-flow-mlflow --version 1.0.0 > values.yaml
vim values.yaml
# Deploy the helm chart
helm install odahu-flow/odahu-flow-mlflow --name odahu-flow-mlflow --namespace odahu-flow --debug -f values.yaml --atomic --wait --timeout 120
```

Warning

Odahu-flow must be deployed before the mlflow trainer.

MLProject file

Let's look at how the MLProject file is related to Model Training API.

```
name: My Project

entry_points:
  main:
    parameters:
      data_file: path
      regularization: {type: float, default: 0.1}
      command: "python train.py -r {regularization} {data_file}"
  test:
    parameters:
      data_file: path
      command: "python validate.py {data_file}"
```

Model Training API can contain only one entry point. You have to add all hyperparameters, which do not have a default value, to a Model Training. Next, you can find the Model Trainings for the MLProject file.

```
spec:
  entrypoint: main
  hyperParameters:
    data_file: test/123.txt
    regularization: 0.2
```

```
spec:
  entrypoint: main
  hyperParameters:
    data_file: test/456.txt
```

```
spec:
  entrypoint: test
  hyperParameters:
    data_file: test/123.txt
```

MLFlow protocol

Odahu-flow requires that a model be logged through [mlflow API](#).

Example of sklearn model logging:

```
mlflow.sklearn.log_model(lr, "model")
```

Optionally, you can provide input and output samples for Odahu-flow. It allows determining input and output types for Odahu-flow packagers. These names must be *head_input.pkl* and *head_output.pkl*, respectively.

Example of input and output samples logging:

```
train_x.head().to_pickle('head_input.pkl')
mlflow.log_artifact('head_input.pkl', 'model')
train_y.head().to_pickle('head_output.pkl')
mlflow.log_artifact('head_output.pkl', 'model')
```

Model Packagers

Odahu-flow packagers turn a **Trained Model Binary** artifact into a specific application. You can find the list of out-of-the-box packagers below:

- Docker REST
- Docker CLI

Installation

A packager installation is the creation of a new `PackagingIntegration` entity in the **API service**. The most straightforward way is to deploy the *odahu-flow-packagers* helm chart.

```
# Add the odahu-flow helm repository
helm repo add odahu-flow 'https://raw.githubusercontent.com/odahu/odahu-helm/master/'
helm repo update
# Fill in the values for the chart or leave the default values
helm inspect values odahu-flow/odahu-flow-packagers --version 1.0.0-rc35 > values.yaml
vim values.yaml
# Deploy the helm chart
helm install odahu-flow/odahu-flow-packagers --name odahu-flow-packagers --namespace odahu-flow --debug -f values.yaml --atomic --wait --timeout 120
```

Warning

Odahu-flow must be deployed before the packagers installation.

General packager structure

All packagers have the same structure. But different packagers provide a different set of arguments and targets. You can find the description of all fields below:

Packager API

```
# Unique value among all packagers
id: "id-12345"
spec:
  # Type of a packager. Available values: docker-rest, docker-cli.
  integrationName: docker-rest
  # Training output artifact name
  artifactName: wine-model-123456789.zip
  # Compute resources
  resources:
    limits:
      cpu: 1
      memory: 1Gi
    requests:
      cpu: 1
      memory: 1Gi
  # List of arguments. Depends of a Model Packaging integration.
  # You can find specific values in the sections below.
  # This parameter is used for customizing of a packaging process.
  arguments: {}
  # List of targets. Depends of a Model Packaging integration.
  # You can find specific values in the sections below.
```

Packagers management

```
# A packager can interact with a Docker registry, PyPi repository, and so on.  
# You should provide a list of connections for a packager can get access to them.  
targets: []  
# You can set connection which points to some bucket where the Trained Model Binary is stored  
# then packager will extract your binary from this connection  
outputConnection: custom-connection  
# Every packager saves its results into status field.  
# Example of fields: docker image or python packager name.  
status:  
  results:  
    - name: some_param  
      value: some_value
```

Note

You can find an artifactName in the `status.artifactName` field of a model training entity.

Packagers management

Packagers can be managed using the following ways.

Swagger UI

ModelPackaging and PackagingIntegration are available on the Swagger UI at <http://api-service/swagger/index.html> URL.

Odahu-flow CLI

Odahuflowctl supports the Packagers API. You must be logged in if you want to get access to the API.

Getting all packaging in json format:

```
odahuflowctl pack get --format json
```

Getting the arguments of the packagers:

```
odahuflowctl pack get --id tensorflow-cli -o 'jsonpath=[*].spec.arguments'
```

- Creating of a packager from `pack.yaml` file:

```
odahuflowctl pack create -f pack.yaml
```

- All packager commands and documentation:

```
odahuflowctl pack --help
```

JupyterLab

Odahu-flow provides the JupyterLab extension for interacting with Packagers API.

Docker REST

The Docker REST packager wraps an ML model into the REST service inside a Docker image. The resulting service can be used for point prediction thorough HTTP.

The packager provides the following list of targets:

| Target Name | Connection Types | Required | Description |
|-------------|------------------|----------|---|
| docker-push | docker, ecr | True | The packager will use the connection for pushing a Docker image result |
| docker-pull | docker, ecr | False | The packager will use the connection for pulling a custom base Docker image |

The packager provides the following list of arguments:

| Argument Name | Type | Default | Required | Description |
|---------------|---------|--|----------|--|
| imageName | string | <code>{{ Name }}-{{ Version }}:{{ RandomUUID }}</code> | False | This option provides a way to specify the Docker image name. You can hardcode the full name or specify a template. Available template values: Name (Model Name), Version (Model Version), RandomUUID. Examples: myservice:123, <code>{{ Name }}:{{ Version }}</code> |
| port | integer | 5000 | False | Port to bind |
| timeout | integer | 60 | False | Serving timeout in seconds. |
| workers | integer | 1 | False | Count of serving workers |
| threads | integer | 4 | False | Count of serving threads |
| host | string | 0.0.0.0 | False | Host to bind |

Docker REST

| | | | | |
|--------------------------------|---------|------------------|-------|---|
| dockerfileBaseImage | string | python:3.6 | False | Base image for Dockerfile |
| dockerfileAddCondaInstallation | boolean | True | False | Add conda installation code to Dockerfile |
| dockerfileCondaEnvLocation | boolean | /opt/conda/envs/ | False | Conda env location in Dockerfile |

The packager provides the following list of result fields:

| Name | Type | Description |
|-------|--------|---------------------------------------|
| image | string | The full name of a built Docker image |

Let's build a couple of examples of Docker REST packager. The packager requires docker or ecr connection types. The following example assumes that you have created a connection with `test-docker-registry` id and `gcr.io/project/odahuflow` URI.

Minimal Example of Docker REST packager

```
id: "docker-rest-packager-example"
spec:
  integrationName: docker-rest
  artifactName: wine-model-123456789.zip
  targets:
    - connectionName: test-docker-registry
      name: docker-push
```

Then a result of the packager will be something like this:

`"gcr.io/project/odahuflow/wine-0-1:ec1bf1cd-216d-4f0a-a62f-bf084c79c58c"`.

Now, let's try to change the docker image name and number of workers.

Docker REST packager with custom arguments

```
id: "docker-rest-packager-example"
spec:
  integrationName: docker-rest
  artifactName: wine-model-123456789.zip
  targets:
    - connectionName: test-docker-registry
      name: docker-push
  arguments:
    imageName: "wine-test:prefix-{{ RandomUUID }}"
    workers: 4
```

```
odahuflowctl pack get --id "docker-rest-packager-example" -o 'jsonpath=$[0].status.results[0].value'
```

Then a result of the packager will be something like this:

`"gcr.io/project/odahuflow/wine-test:prefix-ec1bf1cd-216d-4f0a-a62f-bf084c79c58c"`.

You can run the image locally using the following command:

Docker CLI

```
docker run -it --rm --net host gcr.io/project/odahuflow/wine-test:prefix-ec1bf1cd-216d-4f0a-a62f-bf084c79c58c
```

The model server provides two urls:

- GET `/api/model/info` - provides a swagger documentation for a model
- POST `/api/model/invoke` - executes a prediction

```
curl http://localhost:5000/api/model/info
curl -X POST -d '{"columns": ["features", "features", "features"], "data": [[1, 2, 3], [4, 5, 6]]}' -H "Content-Type: application/json" http://localhost:5000/api/model/invoke
```

Docker REST predict API

```
{
  "columns": [
    "features",
    "features",
    "features",
    "features"
  ],
  "data": [
    [
      [
        1,
        2,
        3,
      ],
      [
        4,
        5,
        6,
      ]
    ]
  ]
}
```

Docker REST prediction result

```
{
  "prediction": [
    [
      0.09405578672885895
    ],
    [
      0.01238546592343845
    ]
  ],
  "columns": [
    "predictions"
  ]
}
```

Docker CLI

The Docker CLI packager wraps an ML model into the CLI inside a Docker image. The resulting service can be used for batch prediction.

The packager provides the following list of targets:

| Target Name | Connection Types | Required | Description |
|-------------|------------------|----------|-------------|
|-------------|------------------|----------|-------------|

| | | | |
|-------------|-------------|--|---|
| docker-push | docker, ecr | The packager will use the connection for pushing a Docker image result | |
| docker-pull | docker, ecr | False | The packager will use the connection for pulling a custom base Docker image |

The packager provides the following list of arguments:

| Argument Name | Type | Default | Required | Description |
|---------------------------------|---------|-------------------------------------|----------|---|
| imageName | string | {{Name}}-{{Version}}:{{RandomUUID}} | False | This option provides a way to specify the Docker image name. You can hardcode the full name or specify a template. Available template values: Name (Model Name), Version (Model Version), RandomUUID. Examples: myservice:123, {{Name}}:{{Version}} |
| dockerfileBaseImage | string | python:3.6 | False | Base image for Dockerfile |
| dockerfileAddCondaInstallations | boolean | True | False | Add conda installation code to Dockerfile |

Docker CLI

| | | | | |
|---------------------------|--------|------------------|-------|----------------------------------|
| dockerfileCondEnvLocation | string | /opt/conda/envs/ | False | Conda env location in Dockerfile |
|---------------------------|--------|------------------|-------|----------------------------------|

The packager provides the following list of result fields:

| Name | Type | Description |
|-------|--------|---------------------------------------|
| image | string | The full name of a built Docker image |

Let's build a couple of examples of Docker CLI packager. The packager requires docker or ecr connection types. The following example assumes that you have created a connection with `test-docker-registry` id and `gcr.io/project/odahuflow` URI.

Minimal Example of Docker CLI packager

```
id: "docker-cli-packager-example"
spec:
  integrationName: docker-cli
  artifactName: wine-model-123456789.zip
  targets:
    - connectionName: test-docker-registry
      name: docker-push
```

Then a result of the packager will be something like this:

`"gcr.io/project/odahuflow/wine-0-1:ec1bf1cd-216d-4f0a-a62f-bf084c79c58c"`.

Now, let's try to change the docker image name and the base image.

Docker CLI packager with custom arguments

```
id: "docker-cli-packager-example"
spec:
  integrationName: docker-cli
  artifactName: wine-model-123456789.zip
  targets:
    - connectionName: test-docker-registry
      name: docker-push
  arguments:
    imageName: "wine-test:prefix-{{ RandomUUID }}"
    dockerfileBaseImage: "python:3.7"
```

```
odahuflowctl pack get --id "docker-cli-packager-example" -o 'jsonpath=$[0].status.results[0].value'
```

Then a result of the packager will be something like this:

`"gcr.io/project/odahuflow/wine-test:prefix-ec1bf1cd-216d-4f0a-a62f-bf084c79c58c"`.

You can run the image locally using the following command:

```
docker run -it --rm --net host gcr.io/project/odahuflow/wine-test:prefix-ec1bf1cd-216d-4f0a-a62f-bf084c79c58c --help
```

The model CLI provides two commands:

- `predict` - Make predictions using GPPI model
- `info` - Show model input/output data schema

Docker CLI info command

Docker CLI

```
docker run -it --rm --net host gcr.io/project/odahuflow/wine-test:prefix-eclbf1cd-216d-4f0a-a62f-bf084c79c58c info
```

Docker CLI info command output

```
Input schema:  
{  
    "columns": {  
        "example": [  
            "features",  
            "features",  
            "features",  
        ],  
        "items": {  
            "type": "string"  
        },  
        "type": "array"  
    },  
    "data": {  
        "items": {  
            "items": {  
                "type": "number"  
            },  
            "type": "array"  
        },  
        "type": "array",  
        "example": [  
            [  
                0,  
                0,  
                0,  
            ]  
        ]  
    }  
}  
Output schema:  
{  
    "prediction": {  
        "example": [  
            [  
                0  
            ]  
        ],  
        "items": {  
            "type": "number"  
        },  
        "type": "array"  
    },  
    "columns": {  
        "example": [  
            "predictions"  
        ],  
        "items": {  
            "type": "string"  
        },  
        "type": "array"  
    }  
}
```

Let's make a batch prediction.

Create a predict file

```
mkdir volume  
cat > volume/predicts.json <<EOL  
{  
    "columns": [  
        "features",  
        "features",  
    ]  
}
```

Docker CLI

```
"features",
],
"data": [
  [
    1,
    2,
    3
  ],
  [
    4,
    5,
    6
  ]
]
}
EOL
docker run -it --rm --net -v volume:/volume host gcr.io/project/odahuflow/wine-test:prefix-eclbf1cd-216d-4f0a-a62f-bf084c79c58c predict /volume/predicts.json /volume
```

Result of prediction

```
cat volumes/result.json
{
  "prediction": [
    [
      0.09405578672885895
    ],
    [
      0.01238546592343845
    ]
  ],
  "columns": [
    "predictions"
  ]
}
```

Model Deployments

Odahu-flow Model Deployment API allows deploy of ML models in a Kubernetes cluster. Additionally, it provides the following set of features:

- Feedback loop
- Scale to zero
- Dynamic Model swagger
- Monitoring of Model Deployment

General deployment structure

Deployment API

```
kind: ModelDeployment
# Some unique value among all deployments
id: wine-12345
spec:
  # Model image is required value. Change it
  image: gcr.io/project/test-e2e-wine-1.0:b591c752-43d4-43e0-8392-9a5715b67573
  # If the Docker image is pulled from a private Docker repository then
  # you have to create a Odahu-flow connecton and specify its id here.
  # imagePullConnID: test

  # Compute resources for the deployment job.
  resources:
    limits:
      cpu: 1
      memory: 1Gi
    requests:
      cpu: 1
      memory: 1Gi

  # Minimum number of replicas
  minReplicas: 0
  # Maximum number of replicas
  maxReplicas: 1
```

Model Deployment management

Model Deployments can be managed using the following ways.

Swagger UI

ModelDeployments are available in the Swagger UI at <http://api-service/swagger/index.html> URL.

Odahu-flow CLI

Service Catalog

Odahuflowctl supports the Model Deployment API. You must be logged in if you want to access the API.

Getting all model deployments in json format:

```
odahuflowctl deployment get --format json
```

Getting the model name of the model deployments:

```
odahuflowctl deployment get --id tensorflow-cli -o 'jsonpath=[*].spec.model.name'
```

- Creating of a training from *train.yaml* file:

```
odahuflowctl deployment create -f train.yaml
```

- All model deployments commands and documentation:

```
odahuflowctl deployment --help
```

- All model deployments commands and documentation:

```
odahuflowctl deployment --help
```

- Getting a model deployment information:

```
odahuflowctl model info --md wine
```

- Making a prediction:

```
odahuflowctl model invoke --md wine --file request.json
```

JupyterLab

Odahu-flow provides the JupyterLab extension for interacting with Model Deployments API.

Service Catalog

Service catalog provides a Swagger UI for Model Deployments.

Note

The model must provide input and output samples to appear in the Service Catalog

Service catalog Swagger UI:

The screenshot shows the Swagger UI interface for a service catalog. At the top, there's a navigation bar with the Swagger logo, a link to `./data.json`, and a green "Explore" button. Below the header, the title "Service Catalog 1.0" is displayed, along with a link to `./data.json`. A brief description follows: "Catalog of model services" and "Terms of service Apache 2.0". Underneath, a dropdown menu for "Schemes" is set to "HTTPS". The main content area is titled "model-123". It lists two endpoints: a blue "GET" button for `/model/model-123/api/model/info` and a green "POST" button for `/model/model-123/api/model/invoke`, which is described as "Prediction".

Example of a prediction request:

This screenshot shows a detailed view of the "POST /model/model-123/api/model/invoke Prediction" dialog. The "Parameters" section is expanded, showing a table with a single row for "PredictionParameters". The "object" parameter is highlighted with a red asterisk indicating it's required. The value is a JSON object representing wine quality features:

```
{ "columns": [ "fixed acidity", "volatile acidity", "citric acid", "residual sugar", "chlorides", "free sulfur dioxide", "total sulfur dioxide", "density", "pH", "sulphates", "alcohol" ], "data": [ [ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ] ] }
```

 There are "Edit Value" and "Model" buttons next to the input field. Below the table, a "Cancel" button is visible. A "Parameter content type" dropdown is set to "application/json". At the bottom, there are "Execute" and "Clear" buttons.

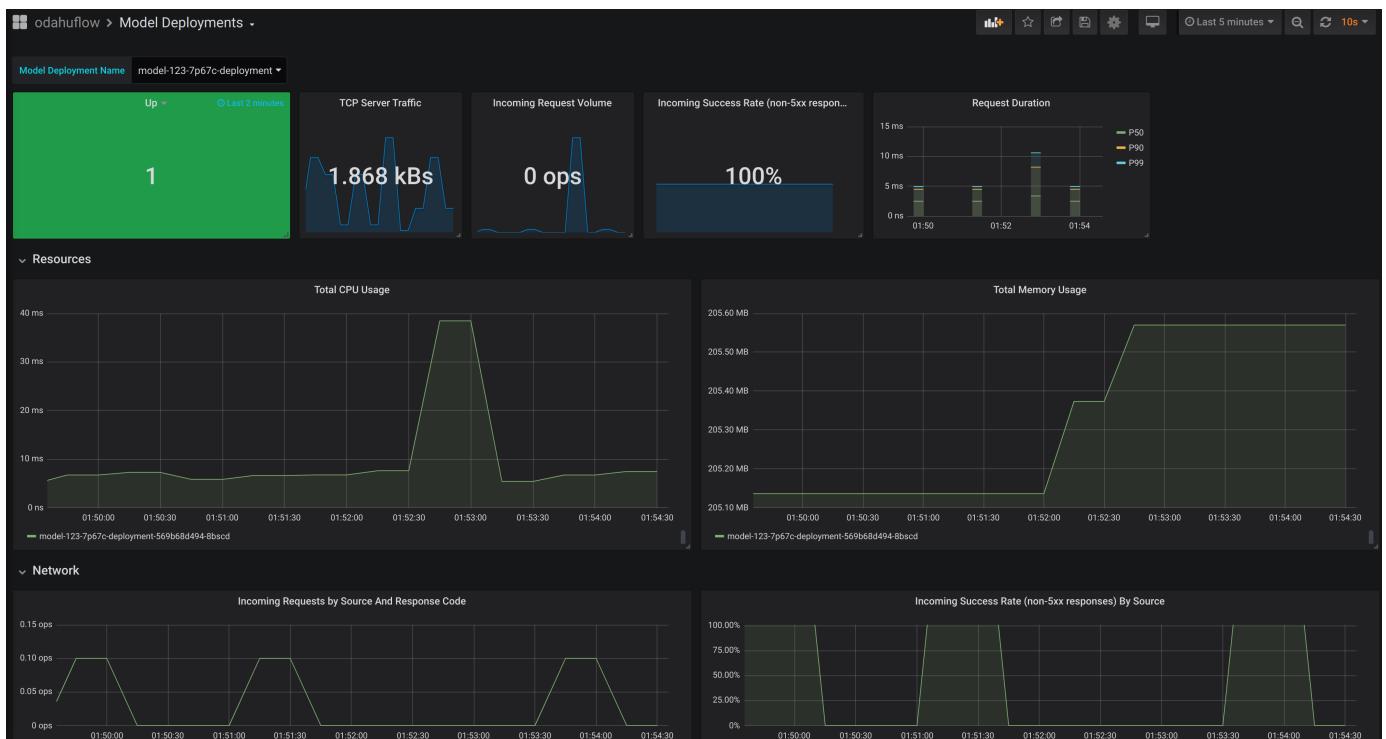
Grafana Dashboard

Out of the box, Odahu-flow provides the Grafana Model Deployment dashboard. It contains the charts with following system metrics:

Feedback

- availability
- replicas
- CPU
- memory
- number of failed HTTP requests
- latency
- ...

Example of the dashboard:



Feedback

Model Feedback provides a view of performance over all stages of model lifecycle.

The mechanism is simple:

1. Ask Deploy for prediction (with or without Request-Id provided)
2. Send prediction feedback to Odahu-flow (with Request-Id returned from previous step)
3. Odahu-flow stores the prediction and feedback to a configurable location

Important

This flow requires feedback to be enabled in `values.yaml` during Helm chart installation

Protocol

1. If prediction is requested without Request-ID: Request-ID header with random ID is added to the request. Otherwise, Request-ID is not generated.
2. Request and response are stored on configured external storage (eg. S3, GCS)
3. User sends Model Feedback as an argument to the feedback endpoint. (Feedback can be arbitrary JSON.)
5. All Feedback is persisted on external storage and can be used by models during subsequent Trains.

Working example

Making a prediction request:

```
curl -X POST -vv "https://cluster-url"/model/"model-deployment-id"/api/model/invoke" \
-H "Authorization: Bearer `JWT`" \
-H "accept: application/json" \
-H "Content-type: application/json" \
-d "{\"columns\": [\"fixed acidity\", \"volatile acidity\", \"citric acid\", \"residual sugar\", \"chlorides\", \"total sulfur dioxide\", \"density\", \"pH\", \"sulphates\", \"alcohol\"], \"data\": [ [ 0, 0, 0, 0, 0, 0, 0, 0, 0 ] ]}"
```

The response contains a generated Request-Id header.

```
HTTP/2 200
server: nginx/1.13.12
date: Tue, 17 Dec 2019 10:58:49 GMT
content-type: application/json
content-length: 45
model-name: test-e2e-wine
model-version: 1.0
request-id: 6fa1f636-fb80-9979-b8c6-d78f5e90f0c1
x-envoy-upstream-service-time: 43
strict-transport-security: max-age=15724800; includeSubDomains

{"prediction": [6.0], "columns": ["quality"]}
```

Requests and responses are persisted in a bucket. (File name `~/request_responses/income/1.1/year=2019/month=07/day=24/2019072414_4.json`)

The first file contains meta-information about request and response.

```
"request_id": "6fa1f636-fb80-9979-b8c6-d78f5e90f0c1",
"model_version": "1.0",
"model_name": "test-e2e-wine",
"response_content": "{\"prediction\": [6.0], \"columns\": [\"quality\"]}",
"request_time": "2019-12-17T08:46:40Z",
"response_time": "2019-12-17T08:46:40Z",
"model_version": "1.0",
"model_name": "test-e2e-wine",
"request_id": "6fa1f636-fb80-9979-b8c6-d78f5e90f0c1",
"request_time": "2019-12-17T08:46:40Z",
"response_time": "2019-12-17T08:46:40Z",
"model_version": "1.0",
"model_name": "test-e2e-wine",
"request_id": "6fa1f636-fb80-9979-b8c6-d78f5e90f0c1",
"request_time": "2019-12-17T08:46:40Z",
"response_time": "2019-12-17T08:46:40Z",
"model_version": "1.0",
"model_name": "test-e2e-wine",
"request_id": "6fa1f636-fb80-9979-b8c6-d78f5e90f0c1",
"request_time": "2019-12-17T08:46:40Z",
"response_time": "2019-12-17T08:46:40Z",
"model_version": "1.0",
"model_name": "test-e2e-wine",
"request_id": "6fa1f636-fb80-9979-b8c6-d78f5e90f0c1",
"request_time": "2019-12-17T08:46:40Z",
"response_time": "2019-12-17T08:46:40Z"}
```

The second file contains the response body with the same Request-ID (File name `~/responses/income/1.1/year=2019/month=07/day=24/2019072414_1.json`)

```
{
  "request_id": "6fa1f636-fb80-9979-b8c6-d78f5e90f0c1",
  "model_version": "1.0",
  "model_name": "test-e2e-wine",
  "response_content": "{\"prediction\": [6.0], \"columns\": [\"quality\"]}",
  "time": "2019-12-17 08:46:40 +0000"
}
```

Working Example - Send Feedback as Payload

Send Model Feedback request:

```
curl -X POST -vv "${BASE_URL}/feedback/model/" \
-H "Authorization: Bearer ${JWT}" \
-H "x-model-name: income" \
-H "x-model-version: 1.1" \
-H "Request-ID: previous-prediction-id" \
-H 'Content-Type: application/json' \
-d '{"truthful": 1}'
```

Note that the -d argument can pass arbitrary JSON.

A successful feedback request will have the following properties:

- HTTP response: 200
- Response field error is false.
- Response field registered is true.
- Response field message is what was sent to storage.

Example response

```
{
  "message": {
    "RequestID": "previous-prediction-id",
    "ModelVersion": "1.0",
    "modelName": "test-e2e-wine",
    "Payload": {
      "json": {
        "truthful": 1
      }
    }
  }
}
```

File name ~=

/feedback/test-e2e-wine/1.0/year=2019/month=11/day=23/2019072311_2.json
will have a format like this, with feedback stored in the payload field:

```
{
  "request_id": "previous-prediction-id",
  "model_version": "1.0",
  "model_name": "test-e2e-wine",
  "payload": {
    "json": {
      "truthful": 1.0
    }
  },
  "time": "2019-12-17 20:08:05 +0000"
}
```

Glossary

VCS

Version control system. A service that stores model source code for development and deployment procedures (e.g. a GitHub Repository).

Trained Model Binary

An archive containing a trained ML/AI model (inference code, model weights, etc). Odahu defines a format for these binaries. See <ref_model_format.html>

Trainer

Application that uses model source code, **Data Bindings, Connections** and **Training Hyperparameters** to produce a **Trained Model Binary**.

Data Binding

Reference to remote data (e.g. files from S3) should be placed for a **Train** process.

Connection

Credentials for an external system. For example: Docker Registry, cloud storage location, etc.

Training Hyperparameters

Parameter for Training process. For example, count of epochs in evolution algorithms.

Train

A containerized process that converts model source code, **Data Bindings, Connections** and **Training Hyperparameters** to **Trained Model Binary** using a **Trainer** defined in a **Trainer Extension**

Trainer Extension

A pluggable **Train** implementation.

Packager

Containerized application that uses a **Trained Model Binary** and **Connections** and converts them into a target Archive. Typically this is a Docker image with REST API.

Package

Containerized process which turns a **Trained Model Binary** into a Docker image with REST API using a **Packager Extension**.

Packager Extension

A pluggable **Package** implementation.

Deployer

Containerized application that uses the results of a **Package** process and **Connections** to deploy a packaged model on a Kubernetes cluster.

Deploy

Containerized process that *Deploys* <Deploy>`results of a :term:`Package` operation to Kubernetes cluster with a REST web service.

Trainer Metrics

Glossary

Metrics set by *Trainer* code during *Train* (e.g. accuracy of model). These metrics can be used for querying and comparing *Train* events.

Trainer Tags

Key/value pairs that are set by *Trainer* code (e.g. type of algorithm). Can be used for querying and comparing *Train* runs.

General Python Prediction Interface

Format of storing models, written in a Python language

MLflow Trainer

Integration of MLflow library for training models, written in a Python. Details - [MLFlow Trainer](#)

REST API Packager

Integration for packing trained models into Docker Image with served via REST API

API service

API for managing Odahu Platform resources for cloud deployed Platform

Operator

A Kubernetes Operator that manages Kubernetes resources (Pods, Services and etc.) for Odahu [Train](#), [Package](#), and [Deploy](#) instances.

Prediction

A deployed model output, given input parameters.

Model prediction API

API provided by deployed models to allow users to request predictions through a web service.

Prediction Feedback

Feedback versus the previous [prediction](#), e.g. prediction correctness.

Model Feedback API

An API for gathering [Prediction Feedback](#)

Feedback aggregator

A service that provides a [Model Feedback API](#) and gathers input and output [prediction requests](#)

Odahu-flow SDK

An extensible Python client library for [API service](#), written in Python language. Can be installed from PyPi.

Odahu-flow CLI

Command Line Interface for [API service](#), written in Python. Can be installed from PyPi. It uses the [Odahu-flow SDK](#).

Plugin for JupyterLab

A odahu-specific plugin that provides Odahu Platform management controls in JupyterLab.

Plugin for Jenkins

A library for managing Odahu Platform resources from Jenkins Pipelines.

Plugin for Airflow

Glossary

A library that provides Hooks and Operators for managing Odahu Platform resources from Airflow.

Model Deployment Access Role Name

Name of scope or role for accessing model deployments.

JWT Token

A JSON Web Token that allows users to query deployed models and to provide feedback. This token contains an encoded **role name**.

A/B testing

Process of splitting predictions between multiple **Model Deployments** in order to compare prediction metrics and **Model Feedback** for models, which can vary by **source code**, **dataset** and/or **training hyperparameters**

Odahu distribution

A collection of Docker Images, Python packages, or NPM packages, which are publicly available for installation as a composable Odahu Platform.

Odahu Helm Chart

A YAML definition for Helm that defines a Odahu Platform deployed on a Kubernetes cluster.

Odahu-flow's CRDs

Objects that **API service** creates for actions that require computing resources to be stored. For example: **connections**, **Trains**, etc).

These objects are Kubernetes Custom Resources and are managed by **operator**.

Development

Pre requirements

To participate in developing of Odahu project you have to meet these requirements:

- Linux/macOS operating systems (due to tools used for development)
- Python:
 - Python v3.6
 - Pipenv
- Golang:
 - Golang v1.12
 - Dep
 - golangci-lint
 - Kubebuilder
 - swag
 - gotestsum
- JupyterLab plugin:
 - Typescript
 - Yarn
- Infra:
 - HELM v2.14.0
 - Kubectl v1.13.2
 - Docker v17+
 - Swagger codegen 2.4.7

Useful links

- Python:
 - Mlflow
 - Robot Framework Guide
 - Odahu Airflow Plugins Development
- Golang:
 - Kubebuilder
- JupyterLab plugin:
 - Typescript handbook

Development

- [React documentation](#)
- [JupyterLab plugin Extension Developer Guide](#)
- Infra:
 - [Helm](#)

A repository directory structure

- containers - docker files
- helms - core helm chart
- packages - source code of packages and applications.
- scripts - utility scripts for CI/CD process.

Repositories

odahu/odahu-flow

Core services of Odahu-flow project.

- odahu-flow-cli python package
- odahu-flow-sdk python package
- E2E Odahu tests
- Training, packaging and deployment operator
- API server
- Feedback services

odahu/odahu-trainer

Collection of training extensions:

- mlflow

odahu/odahu-packager

Collection of model packagers:

- docker-rest

odahu/odahu-flow-jupyterlab-plugin

The jupyterlab-plugin that provides UI for Odahu-flow API service.

odahu/odahu-airflow-plugin

An apache airflow plugin for the Odahu Platform.

odahu/odahu-docs

The repository contains Odahu documentation, which is available [here](#).

odahu/odahu-examples

Repositories

Examples of ML models.

odahu/odahu-infra

Docker images and deployments script for third-party services.

Development hints

Set up a development environment

Odahu product contains 5 main development parts:

- Python packages
 - Executes the `make install-all` command to download all dependencies and install Odahu python packages.
 - Verifies that the command finished successfully, for example: `odahuflowctl --version`
 - Main entrypoints:
 - Odahu-flow SDK - `packages/sdk`
 - Odahu-flow CLI - `packages/cli`
- Odahu-flow JupyterLab plugin
 - Workdir is `odahu/jupyterlab-plugin`
 - Executes the `yarn install` command to download all JavaScript dependencies.
 - Executes the `npm run build && jupyter labextension install` command to build the JupyterLab plugin.
 - Starts the JupyterLab server using `jupyter lab` command.
- Golang services:
 - Executes the `dep ensure` command in the `packages/operator` directory to download all dependencies.
 - Executes the `make build-all` command in the `packages/operator` to build all Golang services.
 - Main entrypoints:
 - API Gateway service - `packages/operator/cmd/edi/main.go`
 - Kubernetes operator - `packages/operator/cmd/operator/main.go`
 - AI Trainer - `packages/operator/cmd/trainer/main.go`
 - AI Packager - `packages/operator/cmd/packager/main.go`
 - Service catalog - `packages/operator/cmd/service_catalog/main.go`
- Odahu-flow Mlflow integration
 - Executes the `pip install -e .` command in the `odahu-flow-mlflow` repository.
- Odahu-flow Airflow plugin
 - Executes the `pip install -e .` command in the `odahu-flow-airflow-plugins` repository.

Update dependencies

- ### Update dependencies
- Python. Update dependencies in a Pipfile. Execute `make update-python-deps` command.
 - Golang. Update dependencies in a Gopkg.toml. Execute `dep ensure` command in packages/operator directory.
 - Typescript. Odahu-flow uses the yarn to manipulate the typescript dependencies.

Make changes in API entities

All API entities are located in packages/operator/pkg/api directory.

To generate swagger documentation execute `make generate-all` in packages/operator directory. Important for Mac users: Makefile uses GNU sed tool, but MacOS uses BSD sed by default. They are not fully compatible. So you need install and use GNU sed on your Mac for using Makefile.

After previous action you can update python and typescript clients using the following command: `make generate-clients`.

Actions before a pull request

Make sure you have done the following actions before a pull request:

- for python packages:
 - `make unittest` - Run the python unit tests.
 - `make lint` - Run the python linters.
- for golang services in the packages/operator directory:
 - `make test` - Run the golang unit tests.
 - `make lint` - Run the golang linters.
 - `make build-all` - Compile all golang Odahu-flow services
- for typescript code in the packages/jupyterlab-plugin directory:
 - `yarn lint` - Run the typescript linter.
 - `jlpm run build` - Compile the jupyterlab plugin.

Local Helm deploy

During development, you often have to change the helm chart, to test the changes you can use the following command quickly: `make helm-install`.

Update dependencies

Optionally, you can create the variables helm file and specify it using the HELM_ADDITIONAL_PARAMS Makefile option. You always can download real variables file from a Terraform state.

Integration Testing

Odahu uses the robotframework for an integration testing. System/integration tests located in the following directories:

- packages/robot - a python package with additional Robot libraries. For example: kubernetes, auth_client, feedback, and so on.
- packages/tests/stuff - artifacts for integration testing. For example: pre-trained ML artifacts, test toolchain integrations, and so on.
- packages/tests/e2e - directory with the RobotFramework tests.

Running system/integration tests

We set up robot tests for gke-odahu-flow-test cluster in the example below. *Do not forget change your cluster url and odahu-flow version.*

Export cluster secrets from odahu-flow-profiles directory.

- Clones *internal* odahu-flow-profiles repository. Checkout your or the developer branch.
- Build hiera docker image using the make docker-build-hiera command in odahu-flow-profiles directory.
- You can optionally override the following parameters in .env file:
 - CLUSTER_NAME
 - HIERA_KEYS_DIR
 - SECRET_DIR
 - CLOUD_PROVIDER
 - EXPORT_HIERA_DOCKER_IMAGE
 - ODAHUFLOW_PROFILES_DIR
- Executes make export-hiera command.
- Verify that \${SECRET_DIR}/.cluster_profile.json file was created.

Updates the .env file when you should override a default Makefile option:

```
# Cluster name
CLUSTER_NAME=gke-odahu-flow-test
# Optionnaly, you can provide RobotFramework settings below.
# Additional robot parameters. For example, you can specify tags or variables.
ROBOT_OPTIONS=-e disable
# Robot files
ROBOT_FILES=**/*.robot
```

Afterward, you should set up a Odahu cluster for RobotFramework tests using the make setup-e2e-robot command. You should execute the previous command only once.

Finally, starts the robot tests:

Integration Testing

```
make e2e-robot
```

Indices and tables

- [Index](#)
- [Module Index](#)
- [Search Page](#)

Index

A

A/B testing

API service

C

Connection

D

Data Binding

Deploy

Deployer

DeploymentOperator (built-in class)

DeploymentSensor (built-in class)

F

Feedback aggregator

G

GcpConnectionToOdahuConnectionOperator (built-in class)

General Python Prediction Interface

J

JWT Token

M

MLflow Trainer

Model Deployment Access Role Name

Model Feedback API

Model prediction API

ModelInfoRequestOperator (built-in class)

ModelPredictRequestOperator (built-in class)

O

Odahu distribution

Odahu Helm Chart

Odahu-flow CLI

Odahu-flow SDK

Odahu-flow's CRDs

Operator

P

Package

Packager

Packager Extension

PackagingOperator (built-in class)

PackagingSensor (built-in class)

Plugin for Airflow

Plugin for Jenkins

Plugin for JupyterLab

Prediction

Prediction Feedback

R

REST API Packager

T

Train

Trained Model Binary

Trainer

Trainer Extension

Trainer Metrics

Trainer Tags

Training Hyperparameters

TrainingOperator (built-in class)

TrainingSensor (built-in class)

V

VCS