|  |  |
|---|---|
| **Tags** | **Kedro**, machine learning, machine learning pipelines, Kubeflow, **Azure**, **Airflow and Vetrex** |
| **Meta Description** | The Kedro framework is one of our recommendations in the case of building reliable machine learning pipelines. In this blog post you will find the list of plugins we created to let you run Kedro on Kubeflow, Airflow, Vertex AI or Azure. |
| **Intro KB** | Do you use Kedro for machine learning pipelines? Check out the list of plugins we created to let you run Kedro on Kubeflow, Airflow, Vertex AI or Azure. |
| **Autor** | Marcin Zabłocki |
| **LINK do grafik** |  |

# Running Kedro… everywhere? Machine Learning Pipelines on Kubeflow, Vertex AI, Azure and Airflow

Building reliable machine learning pipelines puts a heavy burden on Data Scientists and Machine Learning engineers. It's fairly easy to kick-off any ML/AI-driven project with a small team in Jupyter Notebook, but the teams realize fairly quickly that the notebook needs to be transformed into some kind of pipeline, not only to be able to repeat the experiments but also to scale the process e.g. to bigger data or to more sophisticated machine learning models. Here is where Machine Learning Operations (MLOps) come into play. Building machine learning pipelines is one of MLOps best practices.
At GetInData, we hear this story many times from our customers. Our go-to solution for the first part of the problem - building a reliable machine learning pipeline - is usually a recommendation of the Kedro framework. The briefest definition of Kedro is:

*Kedro is an open-source Python framework for creating reproducible, maintainable and modular data science code.*
*Source: Kedro documentation*

On the one hand that's “just” it, on the other, it's really a lot. Kedro provides code structure for machine learning projects by following best software engineering practices without limiting the Data Scientists' capabilities - they can use all of their favorite ML frameworks such as XGBoost, Scikit-Learn, Tensorflow or PyTorch. It introduces abstractions of a pipeline (which describes the logic of the ml project) made of nodes (to perform actions). Nodes consume and output artifacts like data or models (with the abstraction of Data Catalog). It's all enriched with software engineering goodies such as: convention based folder structure, environment aware configuration, event hooks, parametrization, pipeline composition, execution management, unit tests and others. Kedro itself is part of the LF AI & Data foundation and was produced by McKinsey and QuantumBlack in August 2021.

Getting back to the story - by using Kedro we address the following challenges in ML projects:

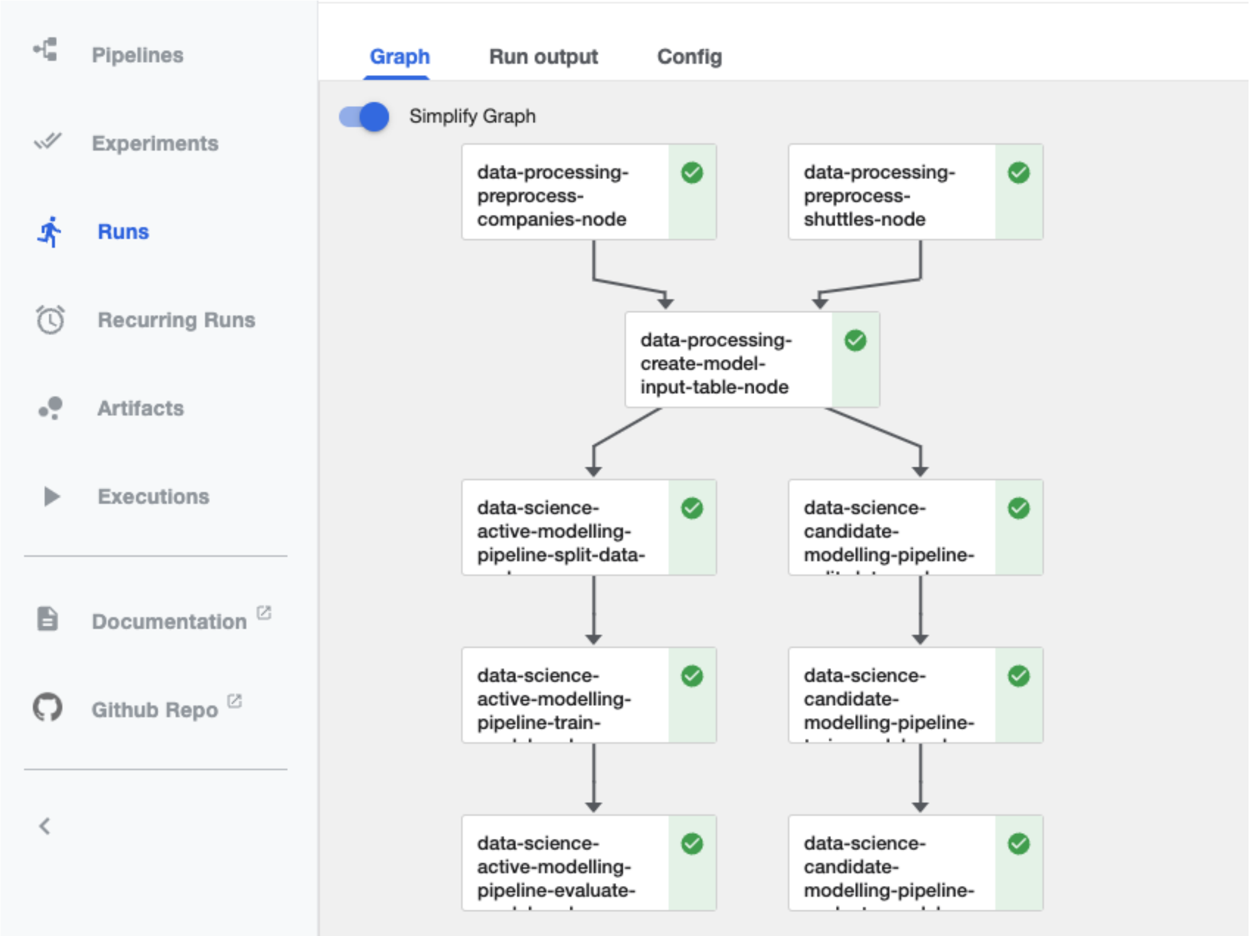
- Pipeline building
- Configuration
- Parametrization
- Datasets management
- Reproducibility
- Data processing
- Model training / evaluation

What about scalability? By default, Kedro runs the pipelines locally (or on a remote machine) either sequentially or in parallel (with the use of Python's multiprocessing, which introduces some challenges). Out-of-the-box, it does not scale well to larger models/datasets as it's limited by the capabilities of the single node (laptop/virtual machine/other).
At GetInData we've solved this limitation by implementing both cloud-native and cloud-agnositc plugins for Kedro, that allow us to run the Kedro pipelines at scale on various serverless/managed cloud services from ALL of the major cloud providers - Google Cloud Platform, AWS and Azure as well as on existing Kubernetes-based infrastructures - Kubeflow Pipelines or Apache Airflow.

## Running Kedro on Kubeflow Pipelines

GitHub: https://github.com/getindata/kedro-kubeflow
Documentation: https://kedro-kubeflow.readthedocs.io/en/stable/

By using the Kedro-Kubeflow plug-in, we enable users to automatically translate Kedro pipelines into Kubeflow Pipelines and run them on KFP clusters. All of the nodes can have distinct CPU, memory or GPU requirements, which enable ML Engineers to easily scale the pipeline in computationally-heavy places, while light nodes might run on lower resources.
As Kubeflow Pipelines itself is cloud-agnostic, you will be able to run and scale your Kedro pipelines in any Kubernetes cluster - either in-cloud (AWS, GCP, Azure) or on-prem.


Pipelines
Experiments
Runs
Recurring Runs
Artifacts
Executions
Documentation
Github Repo
Graph
Run output
Config
Simplify Graph
data-processing-preprocess-companies-node
data-processing-preprocess-shuttles-node
data-processing-create-model-input-table-node
data-science-active-modelling-pipeline-split-data-
data-science-candidate-modelling-pipeline-
data-science-active-modelling-pipeline-train-
data-science-candidate-modelling-pipeline-
data-science-active-modelling-pipeline-evaluate-
data-science-candidate-modelling-pipeline-

*Kedro Spaceflights Tutorial running on Kubeflow Pipelines*
*Kasia - ALT "machine learning pipelines kedro kubflow plugin getindata"*

## Running Kedro on Airflow (in Kubernetes)

GitHub: https://github.com/getindata/kedro-airflow-k8s
Documentation: https://kedro-airflow-k8s.readthedocs.io/en/0.8.0/

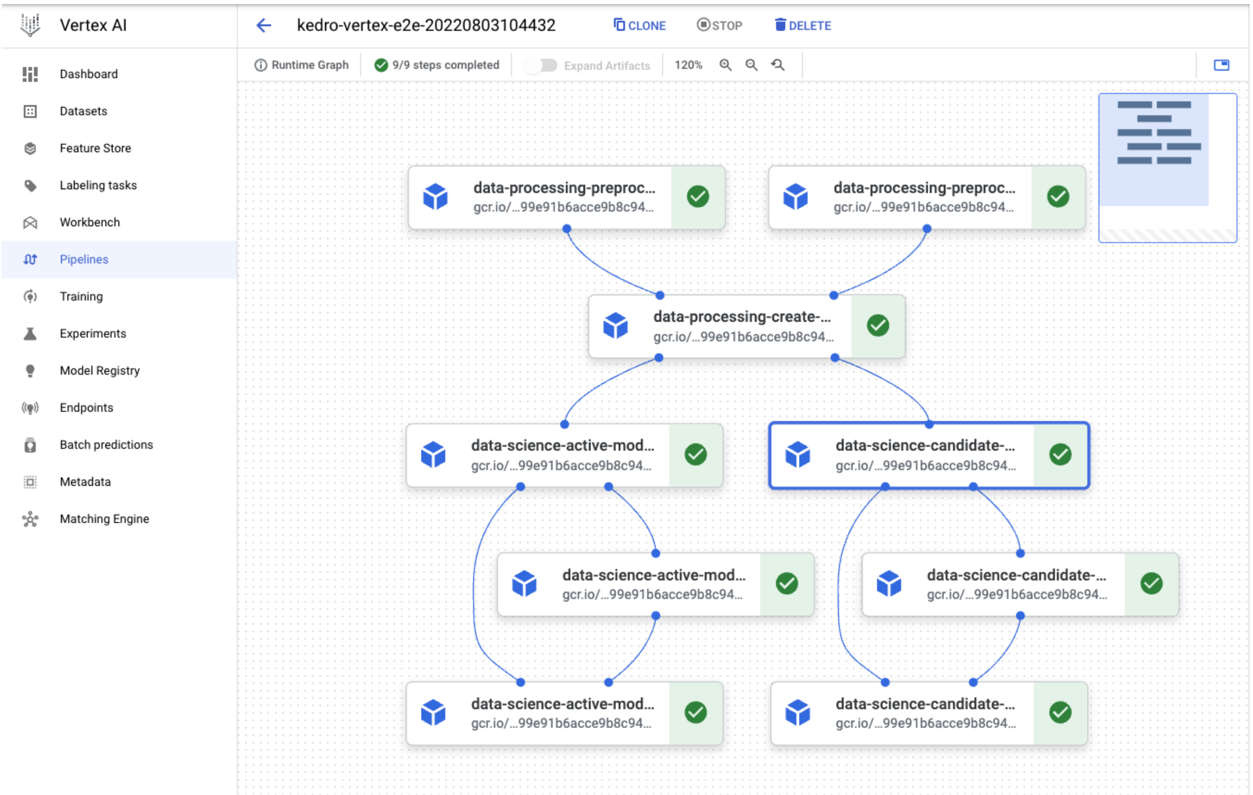
This plugin allows the running of Kedro pipelines with Apache Airflow on Kubernetes Cluster. In contrast to *kedro-airflow,* this plugin does not require additional libraries installed in Airflow runtime, it uses kubernetes infrastructure instead. The plugin itself compiles a Kedro pipeline into Apache Airflow DAG definition - once done, the DAG can be uploaded to an Airflow instance and run. Additionally, the plug-in supports using an external Spark cluster as a computation environment. If you don't want to host Kubernetes, but you like the containerized execution model on AWS, we've got you covered here too - an experimental version of the plugin can run on AWS MWAA (managed Airflow) and deploy containers to ECS instead.

You can read more about kedro-kubeflow and kedro-airflow plugins in our previous blog post Running Machine Learning Pipelines with Kedro, Kubeflow and Airflow written by Mariusz Strzelecki

## Running Kedro on Vertex AI Pipelines

GitHub: https://github.com/getindata/kedro-vertexai
Documentation: https://kedro-vertexai.readthedocs.io/en/stable/

Vertex AI Pipelines is a Google Cloud Platform service that aims to deliver Kubeflow Pipelines functionality in a fully serverless fashion. No manual configuration is needed (and there is no Kubernetes cluster here to maintain - at least not visible to the user). You schedule the pipeline and the service takes care of all of the infrastructure for you.
By using our Kedro-VertexAI plugin, you can run Kedro Pipeline on Vertex AI Pipelines service - again - we do all of the heavy lifting for you. Specify resource requirements (CPU/Memory/GPU) and run the pipeline. It's one of the easiest ways of scaling up Kedro pipelines if you're using the Google Cloud Platform.

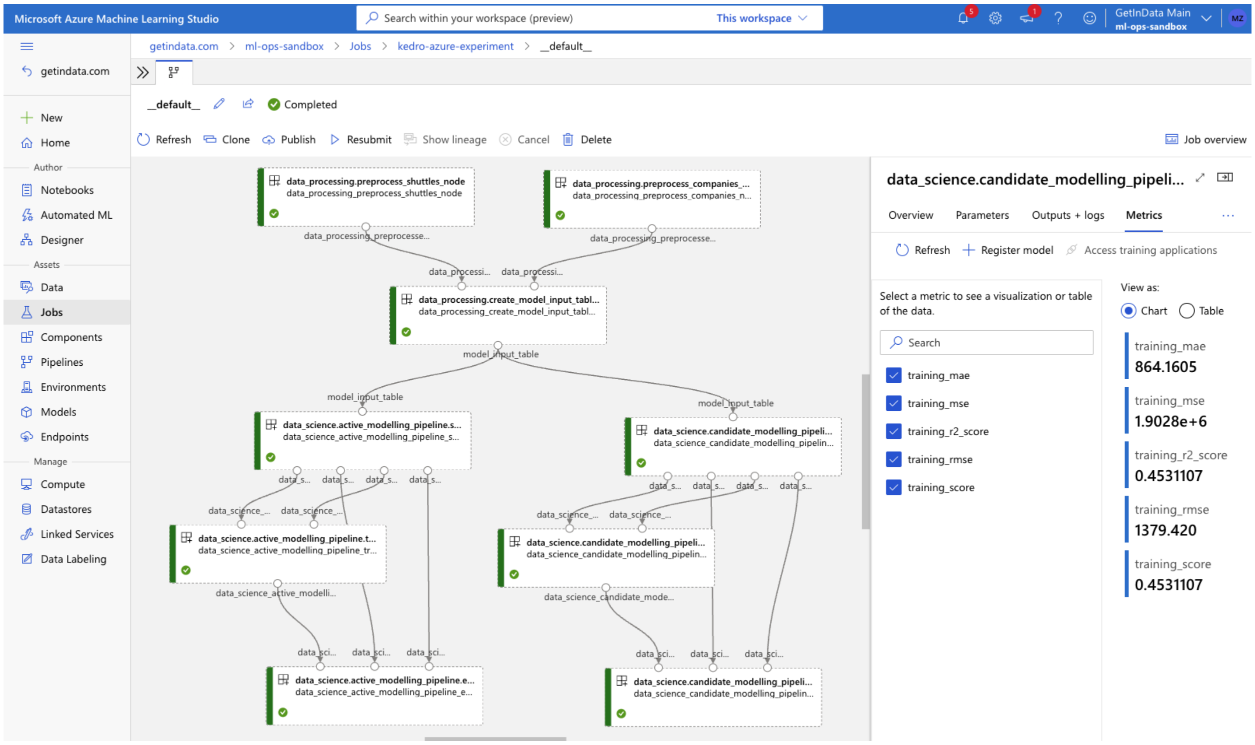

Vertex AI
Dashboard
Datasets
Feature Store
Labeling tasks
Workbench
Pipelines
Training
Experiments
Model Registry
Endpoints
Batch predictions
Metadata
Matching Engine
kedro-vertex-e2e-20220803104432
CLONE
STOP
DELETE
Runtime Graph
9/9 steps completed
Expand Artifacts
120%
data-processing-preproc...
gcr.io/...99e91b6acce9b8c94...
data-processing-preproc...
gcr.io/...99e91b6acce9b8c94...
data-processing-create-...
gcr.io/...99e91b6acce9b8c94...
data-science-active-mod...
gcr.io/...99e91b6acce9b8c94...
data-science-candidate-...
gcr.io/...99e91b6acce9b8c94...
data-science-active-mod...
gcr.io/...99e91b6acce9b8c94...
data-science-candidate-...
gcr.io/...99e91b6acce9b8c94...
data-science-active-mod...
gcr.io/...99e91b6acce9b8c94...
data-science-candidate-...
gcr.io/...99e91b6acce9b8c94...

*Kedro Spaceflights Tutorial running on Vertex AI Pipelines*
*Kasia - ALT "machine learning pipelines kedro vertex ai plugin getindata"*

## Running Kedro on Azure Machine Learning Pipelines

GitHub: https://github.com/getindata/kedro-azureml
Documentation: https://kedro-azureml.readthedocs.io/en/stable/

If your cloud platform of choice is Azure, we've also got your back! Just recently (August 2022) we published the Kedro-AzureML plugin that enables you to run Kedro pipelines in the Azure Machine Learning Pipelines service. Similar to Vertex AI - Azure ML Pipelines is also a fully managed service that provides easy scale-up capabilities. ML Engineers need to set-up compute clusters first, but it's an easy process, which can be handled without the involvement of the DevOps teams. Once done, Kedro pipelines can be executed on Azure ML Pipelines. Moreover, metrics and trained models can be easily tracked using Azure ML's built-in MLflow integration, without any additional configuration and our plug-in supports that.


Microsoft Azure Machine Learning Studio
Search within your workspace (preview)
This workspace
GetInData Main
ml-ops-sandbox
getindata.com
New
Home
Author
Notebooks
Automated ML
Designer
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Compute
Datastores
Linked Services
Data Labeling
getindata.com > ml-ops-sandbox > Jobs > kedro-azure-experiment > _default_
_default_
Completed
Refresh
Clone
Publish
Resubmit
Show lineage
Cancel
Delete
Job overview
data_processing.preprocess_shuttles_node
data_processing.preprocess_companies_...
data_processing.create_model_input_tabl...
model_input_table
data_science.active_modelling_pipeline.s...
data_science.candidate_modelling_pipeli...
data_science.active_modelling_pipeline.t...
data_science.active_modelling_pipeline.e...
data_science.candidate_modelling_pipeli...
Overview
Parameters
Outputs + logs
Metrics
Register model
Access training applications
Select a metric to see a visualization or table of the data.
Search
View as:
Chart
Table

data_science.candidate_modelling_pipeli...

☑ training_mae
☑ training_mse
☑ training_r2_score
☑ training_rmse
☑ training_score

training_mae 864.1605

training_mse 1.9028e+6

training_r2_score 0.4531107

training_rmse 1379.420

training_score 0.4531107

*Kedro Spaceflights Tutorial running on Azure Machine Learning Pipelines*

*Kasia - ALT "machine learning pipelines kedro azure ml plugin getindata"*

We encourage you to watch this step-by-step tutorial:
https://m.youtube.com/watch?v=w_9RzYpGplY
about the Kedro-AzureML plugin that enables you to run Kedro pipelines on Azure ML Pipelines service.

## Summary

We highly encourage you to try out Kedro in your machine learning projects and leverage one of our open-source plug-ins to run and scale Machine Learning pipelines in the cloud! Feel free to reach out to us on GitHub if you encounter any issues with the plug-ins, we're always happy to help. Happy coding!