KubeCon | CloudNativeCon

Europe 2022

WELCOME TO VALENCIA

KubeCon CloudNativeCon Europe 2022

Scaling Open Source ML: How Wolt Uses K8s To Deliver Great Food to Millions

Stephen Batifol, Wolt & Ed Shee, Seldon

About us

Stephen Batifol
Machine Learning Engineer
Wolt

Ed Shee
Head of Developer Relations
Seldon

WOLT IN A NUTSHELL

10 000+
restaurant partners

18M+
registered users

100 000+
courier partners

23
countries

Machine learning @Wolt

- Supply and Demand forecasting
- Recommender systems
- Logistic optimisation
- Fraud detection
- Situation Monitoring Inbox Prioritisation

The different needs we have at Wolt

Data Access

Simple yet **safe** access to **production** data

Infrastructure Access

Scalable computing power at your fingertips. **GPUs**? We got you!

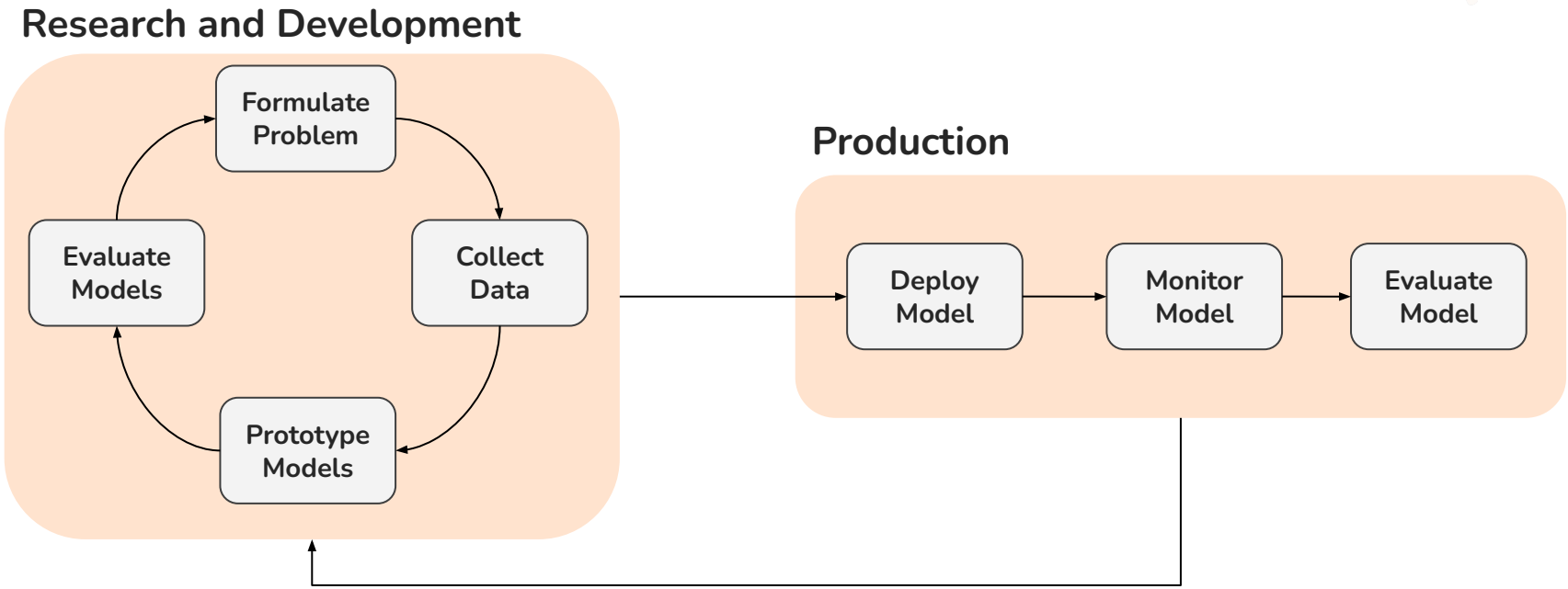
Fast Deployments

Release quicker with the help of **templates** and **CI-CD pipelines**

Standardized Monitoring

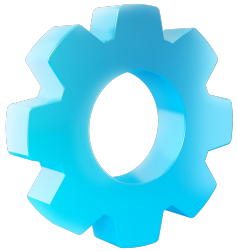
Track data **quality**, training **metrics** and production **performance**

Typical Machine Learning lifecycle



Focus on making iterative work quick and easy

Machine Learning Challenges



Product-thinking

- A few tailored lighthouse projects exist
- New projects are difficult to launch

Tooling

A mix of

- Airflow
- Cronjobs
- Sagemaker
- Custom code

Impact

- Not always known
- Disconnect between business & ML metrics

What do we want for our ML Platform?

Product-thinking

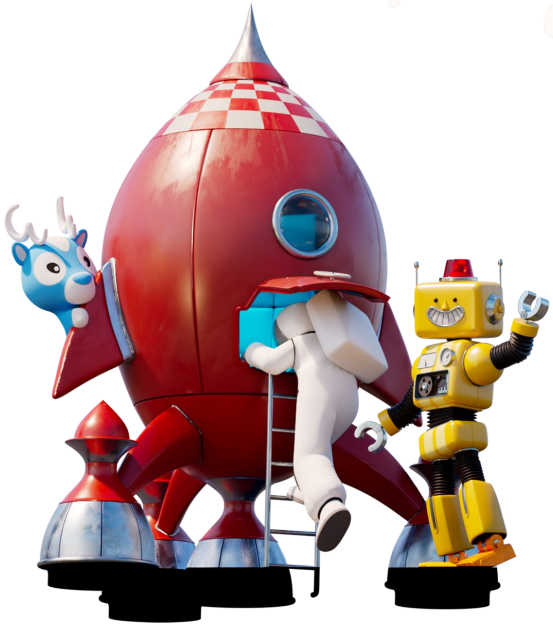
- Easy to launch and iterate
- Driving force for new endeavors
- Focus on platform velocity

Tooling

- Common best-practices
- Flexible and enabling Templates
- Lot of automation

Impact

- Logging and continuous monitoring by default
- Make machine learning is a major differentiator and core business component



HOW DID WE START

CREATE VALUE FOR DATA SCIENTISTS QUICKLY

- **Fast** iteration
- Automatic **monitoring**
- Model **shadowing**
- Rolling update
- ...

EASY TO DEPLOY ON K8S

- Earn the **trust** of the stakeholders and **data scientists**

FOCUS ON ONE COMPONENT

- **Focus** on only **one component** and **make it great.**
- Quick **impact**
- Don't spend too long on things that might not be used

Machine Learning Platform

Train/ Orchestrate Workflows

Distributed, schedulable, scalable - on K8s

Experiment Tracking

Track metrics, store parameters and artifacts, compare experiments

Automatic Update Service

Automatically deploy/ update models

Deployment Service

Models into production microservices, logs and metrics, auto-updates

MLFlow

Projects

Creates a reproducible environment

Defines the model interface and parameters

Models

Standard format for packaging models

Snapshots and versions the model

Tracking

An API to track:

Experiments
Code
Data
Configuration

Model Registry

Centralized model store

Keeps track of model metadata and managed model lifecycle

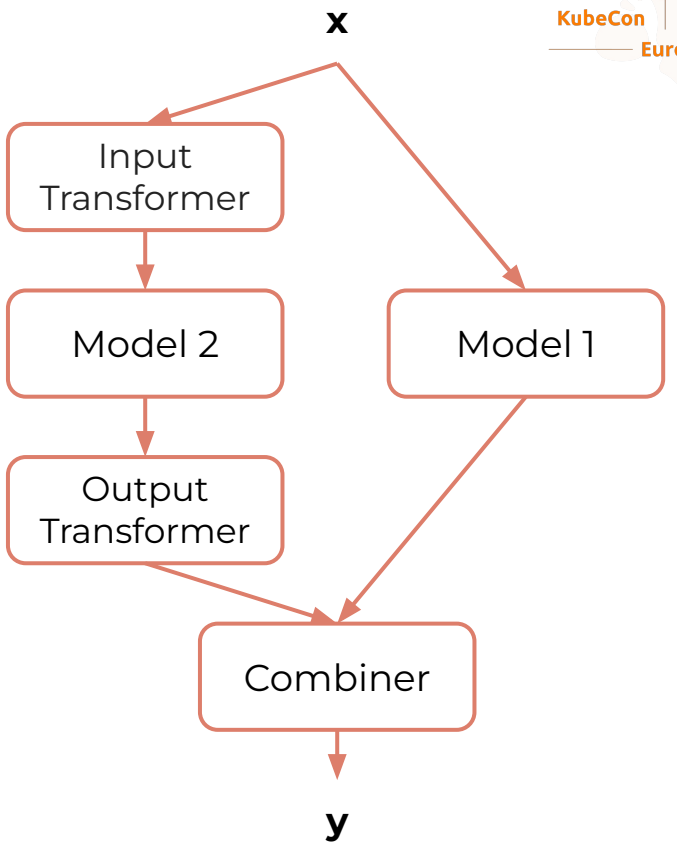
Seldon-core

- Deploys and monitors models on Kubernetes
- Creates containerized microservices with REST and gRPC interfaces
- Provides highly optimized inference servers

```yaml
apiVersion: machinelearning.seldon.io/v1
kind: SeldonDeployment
metadata:
  name: iris-model
  namespace: seldon
spec:
  name: iris
  predictors:
  - graph:
      implementation: SKLEARN_SERVER
      modelUri: gs://seldon-models/v1.14.0-dev/sklearn/iris
      name: classifier
    name: default
    replicas: 1
```

Complex Inference Graphs

```
graph:
  name: node-combiner
  type: COMBINER
  children:
  - name: node-one-input-transformer
    type: TRANSFORMER
    children:
    - name: node-one-model
      type: MODEL
      children:
      - name: node-one-output-transformer
        type: TRANSFORMER
        children: []
  - name: node-two
    type: MODEL
    children: []
```



Integrates with the K8s stack

- Request/response logging to Elasticsearch
- Distributed tracing with Jaeger
- Metrics with Prometheus and Grafana
- Stream processing with Knative and Kafka
- Batch processing with workflow managers (e.g. Argo)

Demo

Food Classification with Tensorflow, MLFlow and Seldon Core

HOW DO WE SCALE OUR ML PLATFORM?

To provide the best experience possible for our Data Scientists we are working on things such as

- **Dedicated K8s clusters**
- **Independent access to resources**
- **Versioned, auditable, and reproducible pipelines**
- **Common tooling and Infrastructure**
- **Automation**
- **Predict future needs of Data Scientists**

EXAMPLE OF A PROBLEM WE FACED

Scale one ML model used by the logistic team.

1. We couldn't really follow how many **Requests Per Second** the model was receiving
2. Nor the **latency**
3. Nor the **errors**
4. We were **hitting the limits** of the scale with the previous infrastructure.
5. Hard to run ad-hoc queries to verify the quality of the model

HOW SELDON-CORE AND K8s HELPED

Seldon-core is running on a K8s cluster

1. Clear idea of how many Requests Per Second it receives (**>400 RPS**)
2. We can **follow the latency and error rates**
3. **Scale up and down** thanks to Horizontal Pod Autoscaling (**HPA**) on CPU and memory usage
4. Rolling **update**
5. **Log** every **response to Kafka** - can now run ad-hoc queries
6. Run **shadow mode**, **A/B tests**

KubeCon CloudNativeCon Europe 2022

Orchestrate and train Machine Learning models in K8s

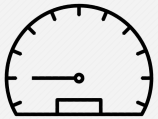
Airflow works great but ...

No support for Dynamic tasks*

Not possible to generate tasks at **Runtime**

Explicit task dependencies

You have to **specify** in which **order** you want your **tasks to run**



No inbuilt caching

Even if your **data is the same**, it will **trigger** the **whole DAG**

Not Designed for ML orchestration

Hard to know which **data** was **used to train** your model, model version, etc.

*Dynamic tasks support added in Airflow 2.3.0 - Released on April 30th 2022

WHY DO WE USE FLYTE?

To solve **some problems** we had with Airflow and is supports **automatic parallelisation**

- **Kubernetes native**
- **Automatic Parallelisation**
- **Versioned, auditable, and reproducible pipelines**
- **Caching**
- **Multi K8s cluster support**
- **Full integration with the rest of our stack**

Future work

- **Advanced** ML model **monitoring**
- Feature engineering for all use cases
- Improve the **ergonomics** of the ML Platform
- **Integration** with the rest of our tools (Experimentation Platform, Data Quality, etc.)

KubeCon CloudNativeCon Europe 2022

Questions?

Stephen Batifol

@_stephenCS

stephen.batifol@wolt.com

/in/stephen-batifol

Ed Shee

@ukcloudman

es@seldon.io

/in/edshee