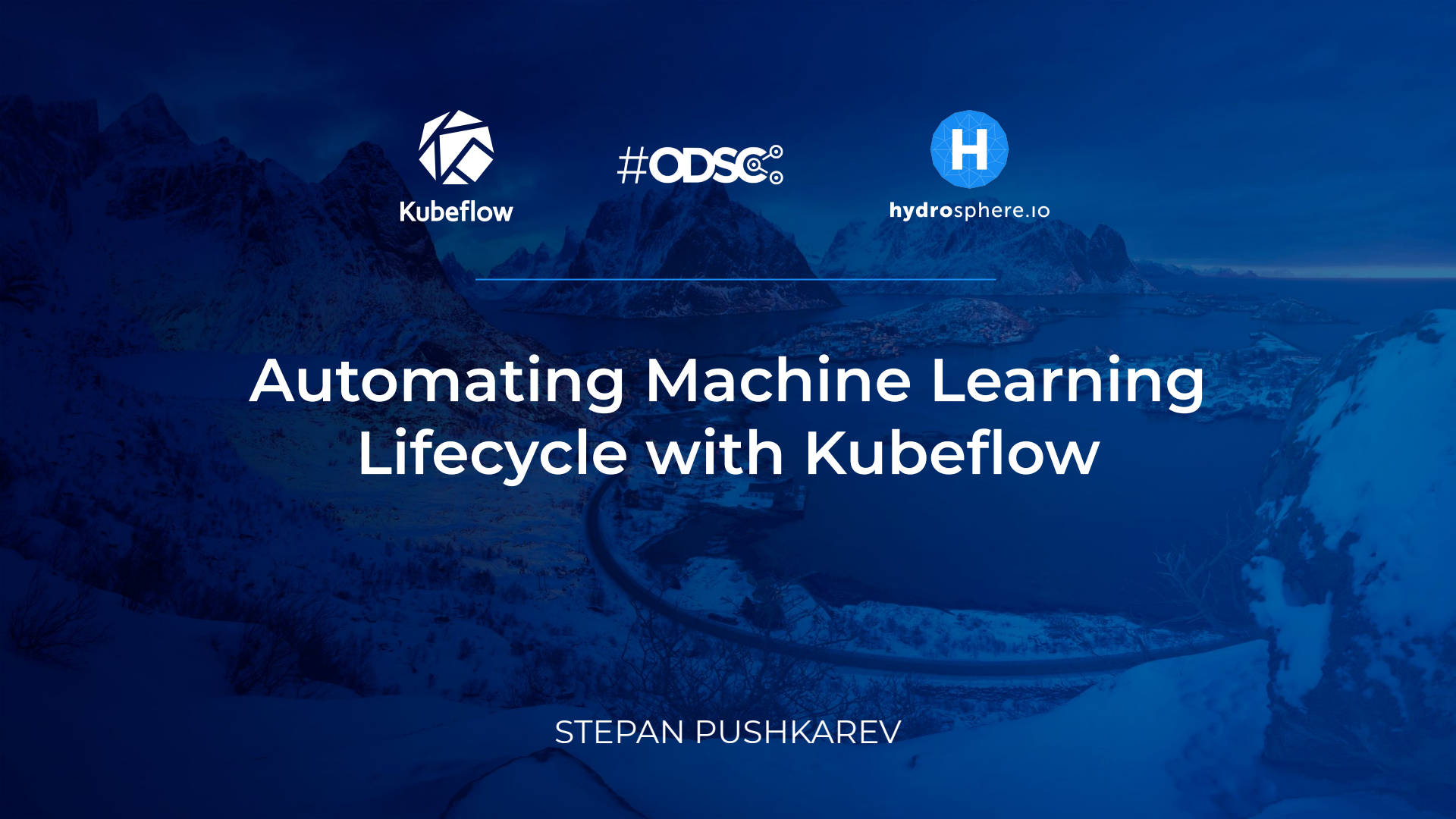


Automating Machine Learning Lifecycle with Kubeflow

STEPAN PUSHKAREV

Outline

1. Intro
2. Why yet another Flow
3. Kubeflow overview

Practice

4. Get a sandbox environment
5. Create pipeline and underlying worker containers
6. Run experiments with Kubeflow

Bonus

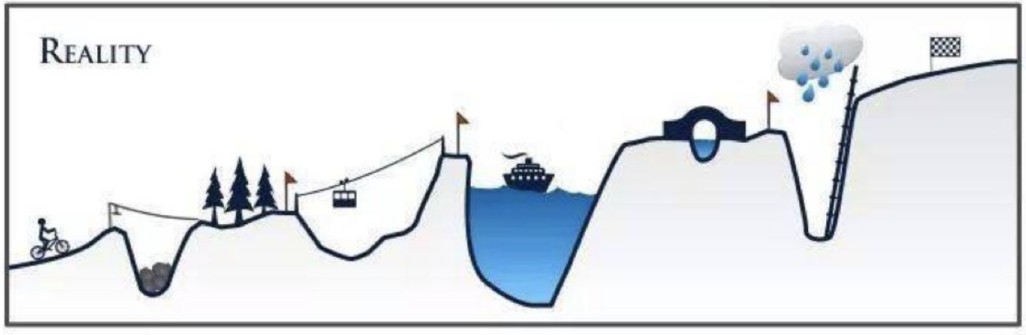
7. Kubeflow vs. MLFlow vs. Airflow

Big Data Science: Expectation vs. Reality

Oct 2016 KDnuggets Gold Blog

By Stepan Pushkarev, CTO, Hydrosphere.io.

The past few years has been like a dream come true for those who work in analytics and big data. There is a new career path for platform engineers to learn Hadoop, Scala and Spark. Java and Python programmers have a chance to move to the Big Data world. There they find higher salaries, new challenges and get to scale up to distributed systems. But recently I am starting to hear some complaints and dashed hopes from engineers who have spent time working there.



Big Data Science: Expectation vs. Reality

1. **Tools evolution**—the Apache Spark/Hadoop ecosystem is great. But it is not stable and user-friendly enough to just run and forget. Engineers and data scientists should contribute to existing opensource projects and create new tools to fill the gaps in day-to-day operations.
2. **Education and cross skills**—when data scientists write code they need to think not just about abstractions but need to consider the practical issues of what is possible and what is reasonable. For example, they need to think how long their query will run and whether the data they extract will fit into the storage mechanism they are using.
3. **Improve the process**—DevOps might be a solution. Here DevOps does not just mean writing Ansible scripts and installing Jenkins. We need DevOps working in optimal fashion to reduce handoff and invent new tools to give everyone self-service to make them as productive as possible.

Why

Machine Learning 5 years ago

Business Problem

Data

High hopes

Then somebody opened a black box....

High hopes

Machine Learning Workflow - whitening the box

1. Research

2. Data Preparation

3. Model Training

4. Model Cataloguing

5. Model Deployment

6. Model Integration Testing

7. Production Inferencing

8. Model Performance Monitoring

9. Model Maintenance

ML Workflow as a pure function

Immutable Raw Dataset

f(x)

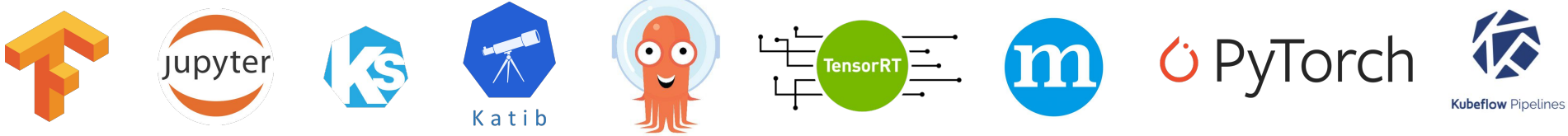
ML Service in prod

Repeatable | Scalable | Observable

Machine Learning Operations

What is Kubeflow?

- Began as Kubernetes template / blueprint for running Tensorflow
- Evolved into “Toolkit” - loosely coupled tools and blueprints for ML on Kubernetes



Kubeflow Pipelines - the first original contribution

Main components:

1. Python SDK
2. UI
3. Orchestrator
4. ML Metadata Service
5. Argo under the hood

Today's Flow Landscape

Data Prep | Training | Cataloguing | Deployment | Integration Testing | Production Inferencing | Performance Monitoring | Model Maintenance

Get sandbox environment

http://odsc.k8s.hydrosphere.io

Workshop modes

1. Intensive - go to github and develop locally
2. Moderate - run through Jupyter and UI
3. TV mode - watch how others do

Local Environment For Intensive mode

1. Clone template repository — https://github.com/Hydrospheredata/odsc-workshop

   $ git clone https://github.com/Hydrospheredata/odsc-workshop.git

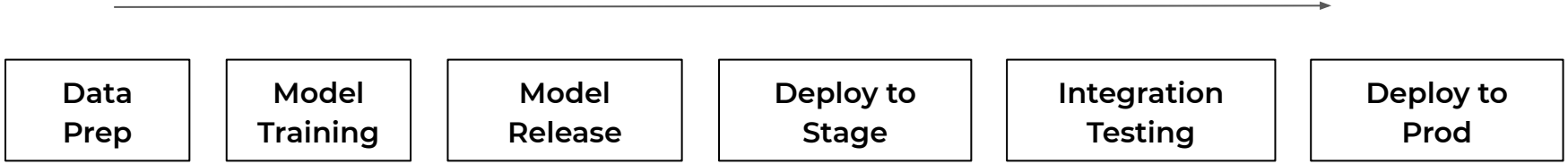
2. Login into Docker account

   $ docker login

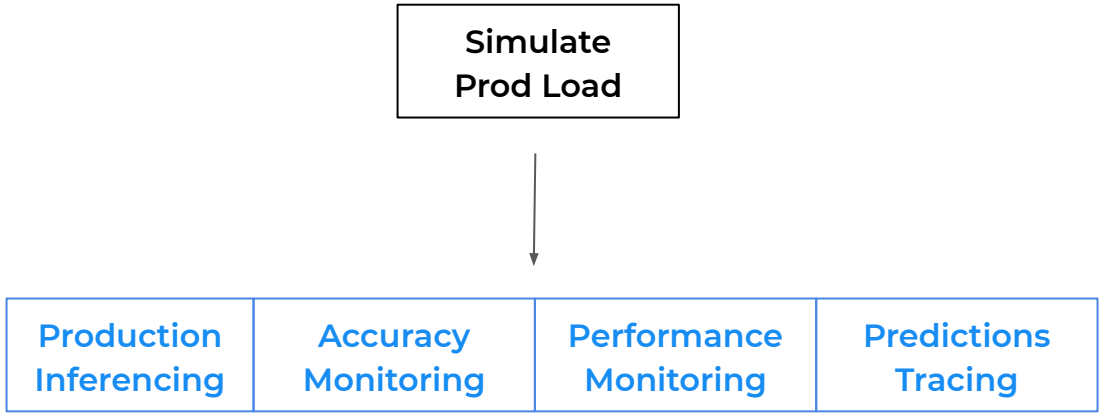
Notes

1. Remember your namespace
2. Do not run more than 2 pipelines in parallel. Terminate the old one and run new.
3. New container launch may take 3-5 minutes. Bare with Kubernetes.

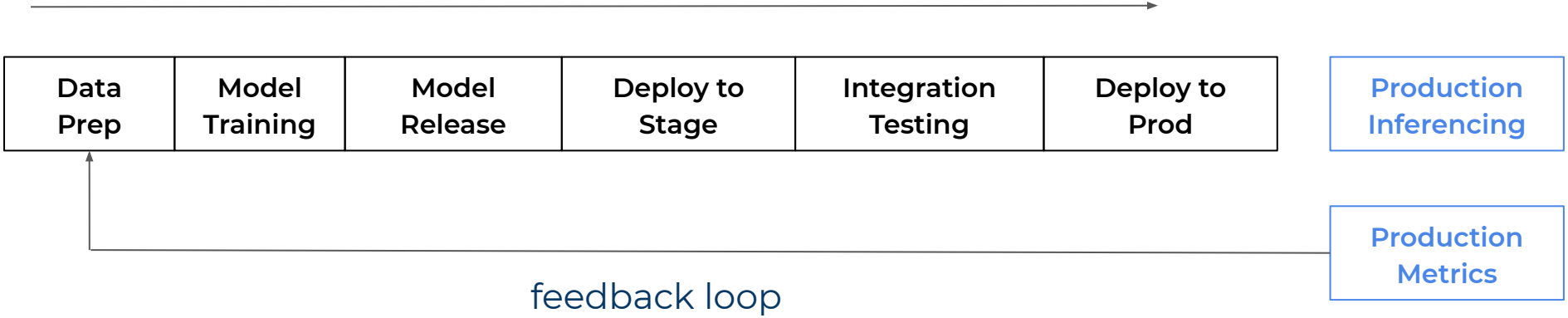
Class plan - Part 1 - Forward path



Class plan - Part 2 - Production Flow



Class plan - Part 3 - Maintenance Flow



Contact Us

GENERAL INQUIRIES

hydrosphere.io
info@hydrosphere.io

linkedin.com/company/hydrospherebigdata
twitter.com/hydrospheredata
facebook.com/hydrosphere.io

BUSINESS AND TECHNICAL

Stepan Pushkarev
spushkarev@hydrosphere.io

Ilnur Garifullin
igarifullin@provectus.com

hydrosphere.io

ADDRESS
125 University Avenue, Suite 290
Palo Alto, CA, 94301
tel: 650-521-7875