AI on Kubernetes

Václav Pavlín
Principal Software Engineer, Office of the CTO
vasek@redhat.com

Agenda

- AI/ML/DS/BD/:)
- State of AI infrastructure
- AI on Kubernetes
- Why is it a good idea?
- Why is it a bas idea?
- Where do I start?
- What’s next?

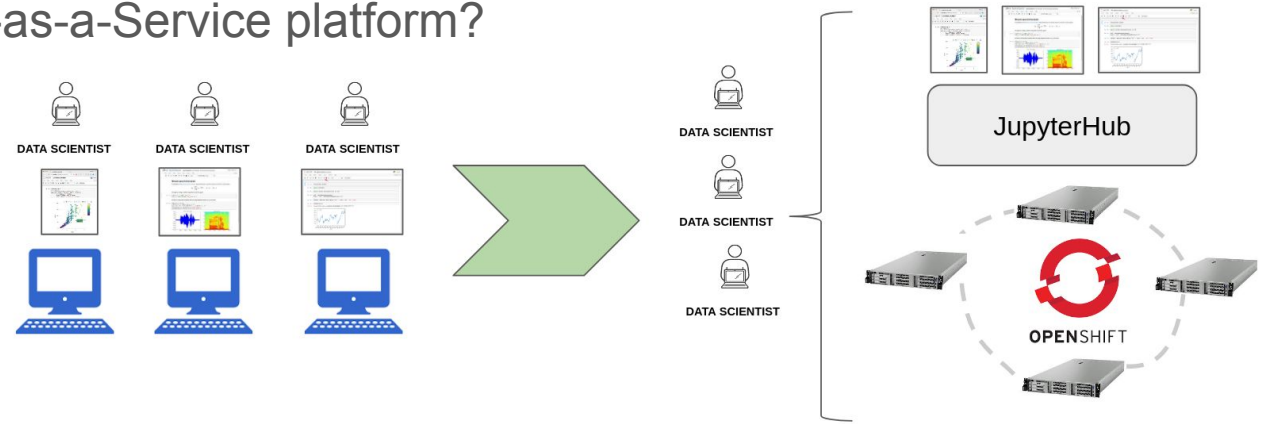
AI/ML/DS/BD/...

- Data is the new oil
- We have
  - ****load of data
  - Reasonable amount of compute resources
- We know
  - Some math
  - Some biology
  - Some programming
- We can let software to sift through the piles of data on its own, kinda...

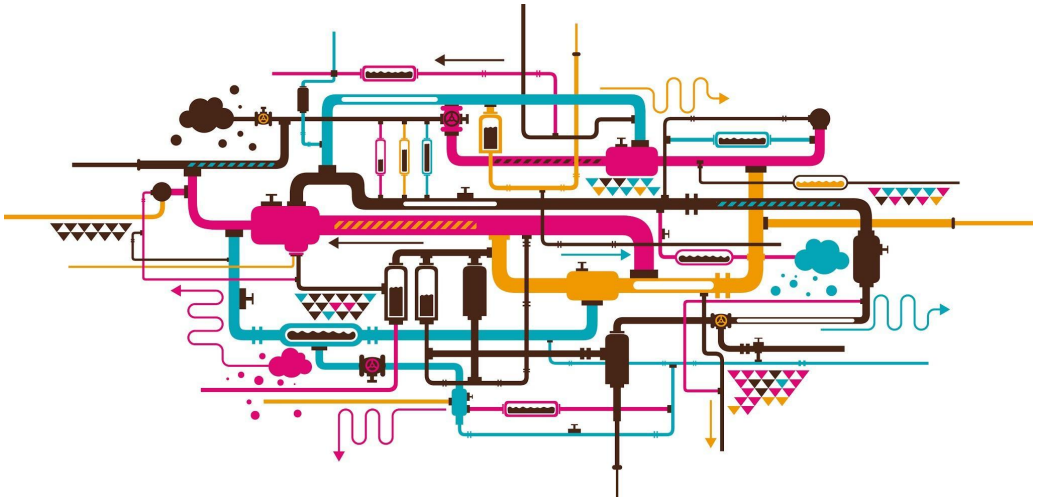
State of AI (infrastructure)

- My laptop is good enough
- I need a workstation
- Can I get a beefy VM?

- Can we get some AI-as-a-Service platform?



AI on Kubernetes



- What do we need?
  - Resources
  - Repeatability
  - Flexibility
  - Multitenancy
  - Storage
- How is this different from “traditional” microservices?

Why is it a good idea?

- Assuming you have Kubernetes cluster available……..
- Kubernetes simplifies
  - Access to resources - compute, storage
  - Experimentation
- Kubernetes promotes
  - Automation - DevOps, GitOps
  - Sharing of resources

Why is it a bad idea?

- When you don’t have a managed Kubernetes cluster:)
- Kubernetes have
  - Pretty steep learning curve (with all the yaml files)
  - Weak solution for data locality
  - Potentially slow networking
- Many AI/ML/DS tools are complicated to setup (for data scientists)

“Come on, make up your mind…
or it’s back to the Infinite Sinkhole of XML.”

Where do I start?

- I do not want to search internet for all the tools and all the options
  - Kubeflow
  - Open Data Hub
  - Some paid solution (IBM Cloud Pak for Data, Cloudera Data Science Workbench, …)

- Minimal set of tools
  - Experiments & Training: Jupyter Notebooks
  - Accessing & Processing big data: Spark
  - Serving models: Seldon
  - Storage: S3 Object Storage (Rook.io Ceph, Minio, ...)

What’s next?

- Kubeflow
  - https://kubeflow.org
  - http://github.com/kubeflow
- Open Data Hub
  - https://opendatahub.io
  - https://gitlab.com/opendatahub/opendatahub-operator/
  - https://github.com/opendatahub-io

Václav Pavlín, vasek@redhat.com, @vpavlin

Questions

- What is your understanding of AI? (Use one sentence)
- Do you run Kubernetes/OpenShift in production?
  - What type of workload do you run there?
- What tools are you interested in from the AI landscape?
- Do you have experience with HPC? How do you feel about transition to Kubernetes?