Kubeflow Explained: Portable Machine Learning on Kubernetes

Michelle Casbon
Strata Data
New York
September 13, 2018
A curated set of compatible tools and artifacts that lays a foundation for running production ML apps

Enables consistency across deployments by providing Kubernetes object templates that bring together disparate components
Agenda

1. Problems
2. Goals
3. What's inside
4. Demo
5. Future Direction
Is this a clearly defined problem?

If not, move along.

If yes, can it be solved in a deterministic way?

If no, dive in.

If yes, do that.

Credit: David Andrzejewski
Counting things is still really hard.
Agenda

1. Problems
2. Goals
3. What's inside
4. Demo
5. Future Direction
Production code
Moving from local to production

Credit: Jörg Wagner and Stefan Prehn
Complexity
Perception

ML Code


@texasmichelle
Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.
Maintainability

- Error resolution, recovery, & prevention
- Speed of iteration
- Versioning
Make it easy for everyone to develop, deploy, and manage portable, scalable ML everywhere
# Kubeflow

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Portability</strong></td>
<td>Entire stack</td>
</tr>
<tr>
<td><strong>Scalability</strong></td>
<td>Native to k8s</td>
</tr>
<tr>
<td></td>
<td>Reduce variability between services &amp; environments</td>
</tr>
<tr>
<td><strong>Composability</strong></td>
<td>Single, unified tool for common processes</td>
</tr>
<tr>
<td><strong>Full product lifecycle</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Support</strong></td>
<td>Specialized hardware, like GPUs &amp; TPUs</td>
</tr>
<tr>
<td></td>
<td>Reduce costs</td>
</tr>
<tr>
<td></td>
<td>Improve model performance</td>
</tr>
</tbody>
</table>
# Kubeflow

<table>
<thead>
<tr>
<th><strong>Who</strong></th>
<th><strong>What</strong></th>
<th><strong>Why</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data scientists</td>
<td>Portable ML products on k8s</td>
<td>Because building a platform is too big of a problem to tackle alone</td>
</tr>
<tr>
<td>ML researchers</td>
<td>v0.2.5 release</td>
<td></td>
</tr>
<tr>
<td>Software engineers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product managers</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

https://github.com/kubeflow/kubeflow
Kubeflow

Kubernetes-native platform for ML
- Run wherever k8s runs
- Use k8s to manage ML tasks
- CRDs for distributed training

Adopt k8s patterns
- Microservices
- Manage infra declaratively

Package infrastructure components together
- Ksonnet
- Move between local -> dev -> test -> prod -> onprem

Support multiple ML frameworks
- Tensorflow
- Pytorch
- Scikit
- Xgboost
- Et al.
Agenda

1. Problems
2. Goals
3. What’s inside
4. Demo
5. Future Direction
But what is it?
A curated set of compatible tools and artifacts that lays a foundation for running production ML apps

Enables consistency across deployments by providing Kubernetes object templates that bring together disparate components
What's Inside?

- Ambassador reverse HTTP proxy
- Central Dashboard
  - JupyterHub
  - Tf-job dashboard
- Tf-job operator
Agenda

1. Problems
2. Goals
3. What's inside
4. Demo
5. Future Direction
Yelp Restaurant Reviews

1. Create a minikube cluster
2. Install Kubeflow locally
3. Run training locally
4. Create a GKE cluster
5. Install Kubeflow on GKE
6. Run training on CPUs
7. Run training on TPUs
8. Create serving and UI
9. Run a notebook on GPUs

https://github.com/kubeflow/examples/demos
Try it Yourself

- codelabs.developers.google.com
  - Intro to Kubeflow on Google Kubernetes Engine: https://goo.gl/192bs7
  - Kubeflow End-to-End: GitHub Issue Summarization: https://goo.gl/qLXUTG
- Qwiklabs: https://qwiklabs.com
- Katacoda: https://www.katacoda.com/kubeflow
- GitHub: https://github.com/kubeflow/examples/tree/master/github_issue_summarization
- http://gh-demo.kubeflow.org
Just the Beginning

- Easier setup
- Utilize more k8s features
- Add support for packages, frameworks, libraries, and example models
- You tell us! Get involved
  - github.com/kubeflow
  - kubeflow.slack.com
  - @kubeflow
  - kubeflow-discuss@googlegroups.com
  - Community call Tuesdays alternating 8:30am and 5:30pm Pacific
  - **Kubeflow Contributor Summit**
    - Sept. 25 in Sunnyvale, CA
Agenda

1. Problems
2. Goals
3. What's inside
4. Demo
5. Future Direction
Future Direction

- Release 0.3.0 end of September
- Getting started: single command
- Monitoring with Prometheus
- Codeless hyperparameter tuning with Katib
- API consistency between TFJob and PyTorch
- TFServing follows k8s style patterns
- Jupyterhub
  - More configurable
  - Run distributed TFJobs from a notebook
- Batch prediction
- Benchmarking
- Chainer support
Questions?