Introduction to Kubeflow Pipelines

Michelle Casbon
Kubecon Europe
Barcelona
May 21, 2019

Google Cloud

TAs

Abhishek Gupta

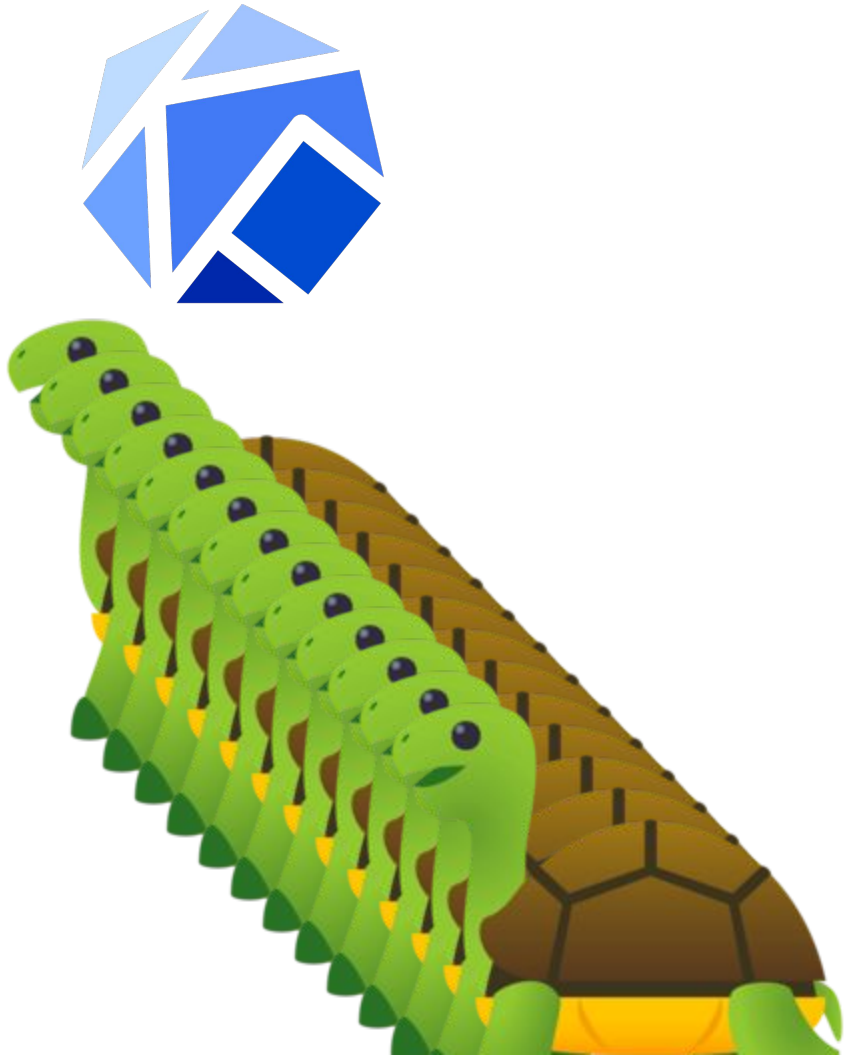
Dan Anghel

Dan Sanche

Jeremy Lewi

Michal Zylinski

Google Cloud



Google Cloud

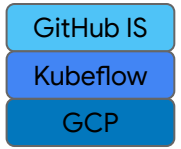
@texasmichelle

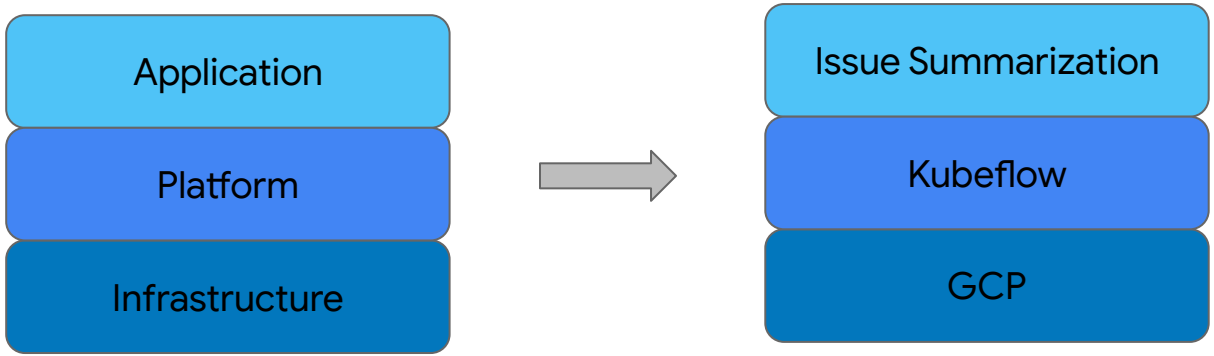
Kubeflow is a curated set of compatible tools and artifacts that lays a foundation for running production ML apps

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

Google Cloud

@texasmichelle





Google Cloud

GitHub IS
Kubeflow
GCP

Contributors

- Kubeflow is open
  - Open community
  - Open design
  - Open source
  - Open to ideas
- 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**
    - Q3 2019

Alibaba Cloud
蚂蚁金服 ANT FINANCIAL
Caicloud
CANONICAL
cisco
DATAWIRE
DELL
GitHub
Google
heptio
HUAWEI
intel
Microsoft
MOMENTA
Pachyderm
jupyter
redhat
Seldon Technologies
UBER
weaveworks

Google Cloud

https://github.com/kubeflow/kubeflow

Agenda

g.co/codelabs/kfp-gis

1. Set up the environment
2. Create a Kubeflow cluster
3. Run a pipeline from the Kubeflow Pipelines dashboard
4. Run a pipeline from a Jupyter notebook
5. Clean up

Google Cloud

Agenda

g.co/codelabs/kfp-gis

Zone: europe-west1-b, europe-west1-d

1. Set up the environment
2. Create a Kubeflow cluster
3. Run a pipeline from the Kubeflow Pipelines dashboard
4. Run a pipeline from a Jupyter notebook
5. Clean up

Google Cloud

PROBLEM

Moving from local to production

SOLUTION

Portability

Package infrastructure components together

GitHub IS
Kubeflow
GCP

Google Cloud

Credit: Jörg Wagner and Stefan Prehn


chrome
backspace
enter
shift
ctrl
alt
tab

PROBLEM

Complexity

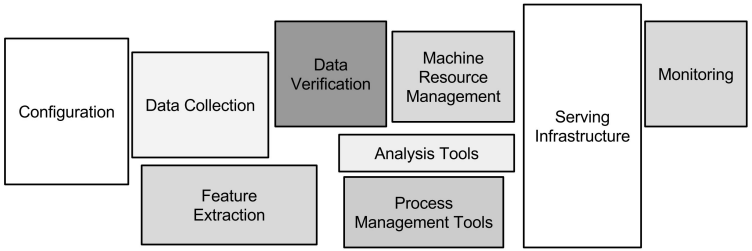
GitHub IS
Kubeflow
GCP

Google Cloud

GitHub IS
Kubeflow
GCP

Perception

ML
Code

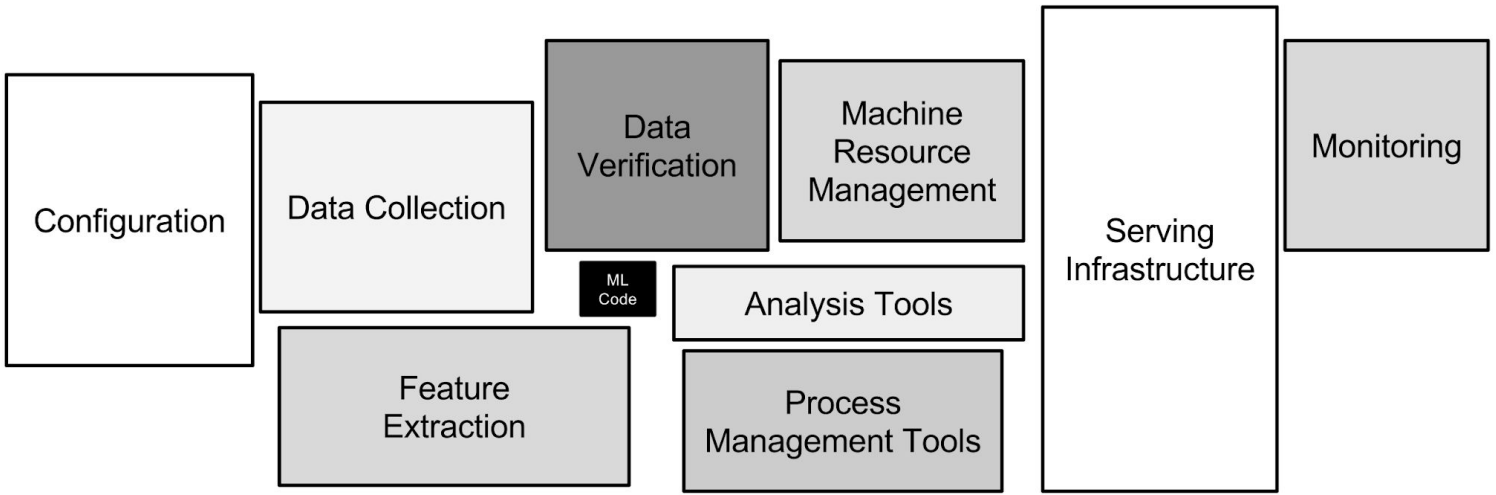


Google Cloud

Credit: Hidden Technical Debt of Machine Learning Systems, D. Sculley, et al.

GitHub IS
Kubeflow
GCP

Reality



Google Cloud

Credit: Hidden Technical Debt of Machine Learning Systems, D. Sculley, et al.

GitHub IS
Kubeflow
GCP

Data

- Data Ingestion
- Data Exploration
- Data Transformation
- Data Validation
- Data Analysis
- Training Data Segmentation

Featurization

- Feature Extraction

Training

- Model Building
- Model Validation
- Model Versioning
- Model Auditing
- Distributed Training
- Continuous Training

Application

- Serving Infrastructure
- Business Logic
- UI
- Load Balancing

Platform

- Configuration
- Process Management
- Resource Management
- Monitoring
- Logging
- Continuous Delivery
- Authentication/ Authorization

Google Cloud

@texasmichelle

PROBLEM

Complexity

SOLUTION

Composability

Logical groupings

Reusable components

GitHub IS

Kubeflow

GCP

Google Cloud

PROBLEM

Maintainability

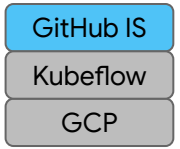
- Error resolution, recovery, & prevention
- Speed of iteration
- Versioning

SOLUTION

Composability

Shorten the development lifecycle

Automation



PROBLEM

Capacity Planning

- Usage patterns
- Demand spikes
- Efficient resource usage

SOLUTION

Scalability

Kubernetes

Autoprovisioning

GitHub IS

Kubeflow

GCP

Google Cloud

Agenda

g.co/codelabs/kfp-gis

Zone: europe-west1-b, europe-west1-d

1. Set up the environment
2. Create a Kubeflow cluster
3. Run a pipeline from the Kubeflow Pipelines dashboard
4. Run a pipeline from a Jupyter notebook
5. Clean up

Google Cloud

Make it easy for everyone to develop, deploy, and manage portable, scalable ML everywhere

Google Cloud

GitHub IS
Kubeflow
GCP

Kubeflow

Portability

Entire stack

Scalability

Native to k8s

Reduce variability between services & environments

Composability

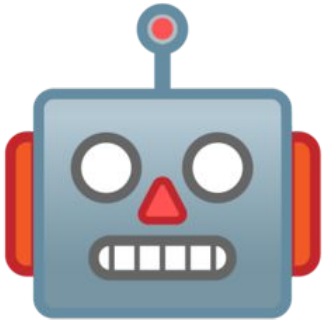
Single, unified tool for common processes

Full product lifecycle

Support specialized hardware, like GPUs & TPUs

Reduce costs

Improve model performance



Google Cloud

GitHub IS
Kubeflow
GCP

Kubeflow

Who

Data scientists

ML researchers

Software engineers

Product managers

What

Portable ML products on k8s

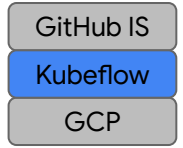
v0.5.0 release

Why

Because building a platform is too big of a problem to tackle alone

`https://github.com/kubeflow/kubeflow`

Google Cloud



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

Kustomize & Ksonnet

Move between local -> dev -> test -> prod -> onprem

Support multiple ML frameworks

TensorFlow

Pytorch

Scikit

Xgboost

Et al.

Google Cloud

Agenda

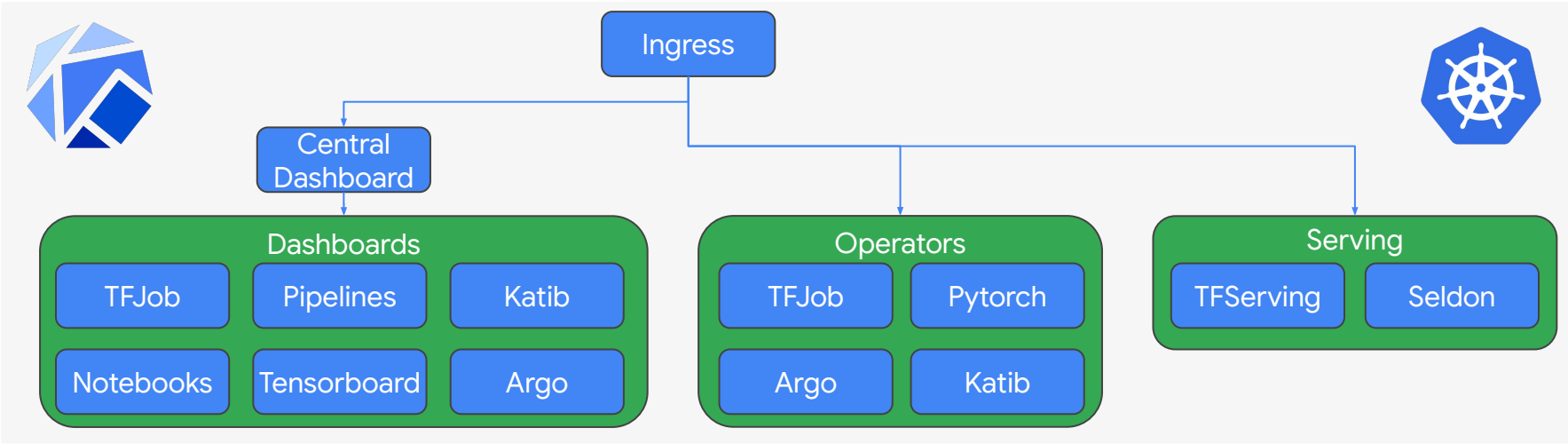
g.co/codelabs/kfp-gis

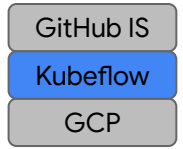
1. Set up the environment
2. Create a Kubeflow cluster
3. Run a pipeline from the Kubeflow Pipelines dashboard
4. Run a pipeline from a Jupyter notebook
5. Clean up

Google Cloud

GitHub IS
Kubeflow
GCP

What's Inside v0.5?





What's new in v0.5?

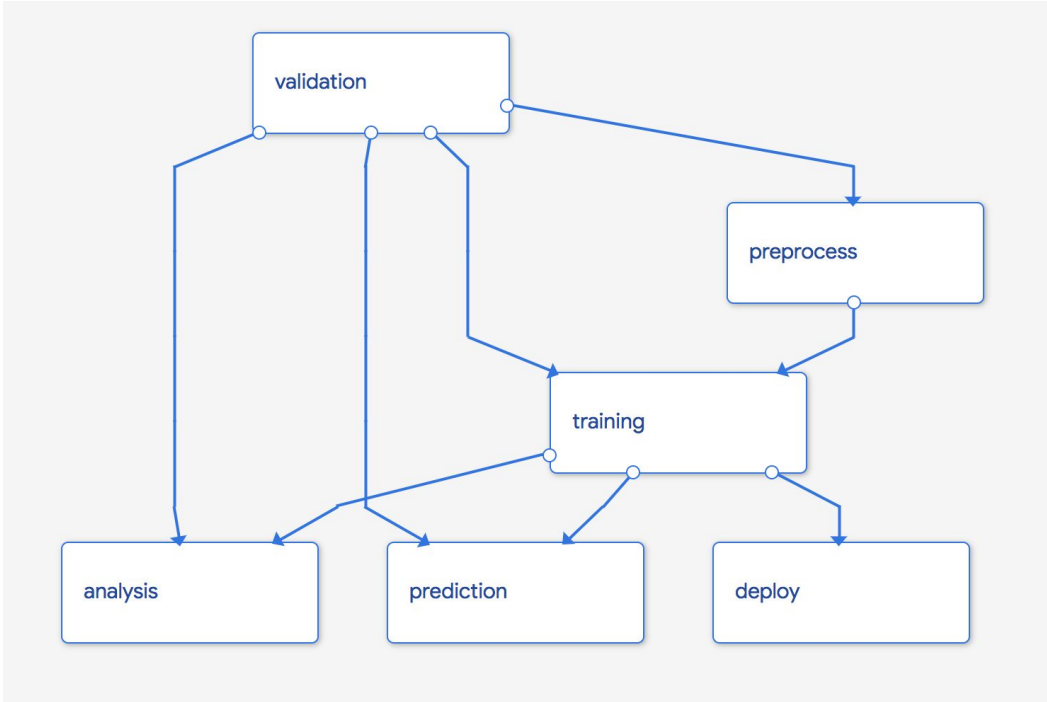
- Deploy
  - CLI: `kfctl` Go binary with bugfixes
  - Click-to-deploy: in-cluster auth
  - Istio support
- Develop
  - Central Dashboard overhaul
  - Notebooks UI
  - Fairing library
    - Build, train, & deploy directly from Python
  - Katib improvements

Google Cloud

GitHub IS
Kubeflow
GCP

Pipelines

- End-to-end ML workflows
- Orchestration
- Service integration
- Components & sharing
- Job tracking, experimentation, monitoring
- Notebook integration



Google Cloud

Agenda

g.co/codelabs/kfp-gis

1. Set up the environment
2. Create a Kubeflow cluster
3. Run a pipeline from the Kubeflow Pipelines dashboard
4. Run a pipeline from a Jupyter notebook
5. Clean up

Google Cloud

Agenda

g.co/codelabs/kfp-gis

1. Set up the environment
2. Create a Kubeflow cluster
3. Run a pipeline from the Kubeflow Pipelines dashboard
4. Run a pipeline from a Jupyter notebook
5. Clean up

Google Cloud

GitHub IS
Kubeflow
GCP

Roadmap

- v0.6 out in early summer
- Multi-user isolation
- Ksonnet replacement
- v1.0 Enterprise readiness
  - Model management
  - Hardened APIs
  - Clean deployments, upgrades
- **You tell us!** (Or better yet, help!)

Google Cloud

GitHub IS
Kubeflow
GCP

Contributors

- Kubeflow is open
  - Open community
  - Open design
  - Open source
  - Open to ideas
- 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**
    - Q3 2019

Alibaba Cloud
蚂蚁金服 ANT FINANCIAL
Caicloud
CANONICAL
cisco
DATAWIRE
DELL
GitHub
Google
heptio
HUAWEI
intel
Microsoft
MOMENTA
Pachyderm
jupyter
redhat
Seldon Technologies
UBER
weaveworks

Google Cloud

https://github.com/kubeflow/kubeflow

Kubeflow Talks @ Kubecon

- **Tutorial Introduction to Pipelines** - *Tuesday May 21 14:00-15:25*; Michelle Casbon, Dan Sanche, Dan Anghel & Michal Zylinski Google (https://sched.co/MPgr)
- **Kubeflow BOF** - *Tuesday May 21 15:55-16:30*; David Aronchick, Microsoft & Yaron Haviv, Iguazio (https://sched.co/PiUF)
- **Building Cross-Cloud ML Pipelines with Kubeflow with Spark & TensorFlow** - *Tuesday May 22 14:00 - 14:35*; Holden Karau, Google & Trevor Grant, IBM (https://sched.co/MPaZ)
- **Toward Kubeflow 1.0, Bringing a Cloud Native Platform for ML to Kubernetes** - *Wed May 22 11:55 - 12:30*; David Aronchick, Microsoft & Jeremy Lewi Google (https://sched.co/MPax)
- **Managing Machine Learning Pipelines In Production with Kubeflow with Devops** - *Wednesday May 22 14:40-15:25* - David Aronchick, Microsoft (https://sched.co/MPaZ)
- **Moving People and Products with Machine Learning on Kubeflow** - Thursday May 23 14:00 -14:35; Jeremy Lewi, Google & Willem Pienaar, GO-JEK (https://sched.co/MPac)

Questions?

Google Cloud

@texasmichelle