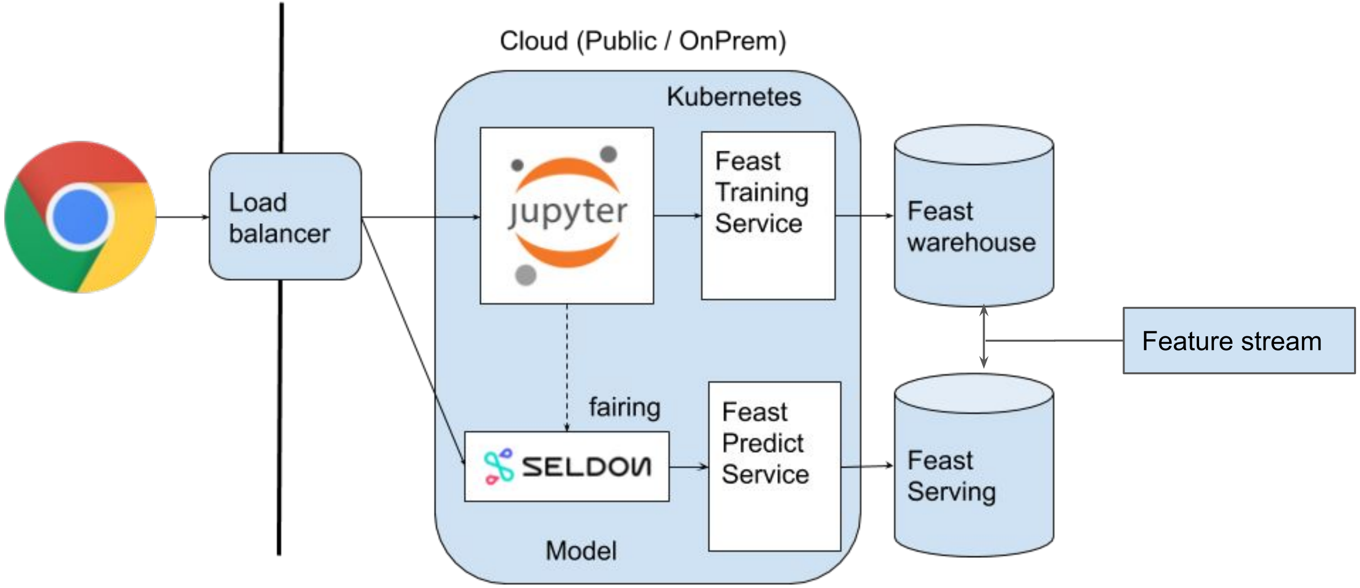
Kubeflow

Moving People and Products with ML on Kubeflow

Jeremy Lewi (jlewi@) Google
Willem Pienaar GOJEK
2019-05-23

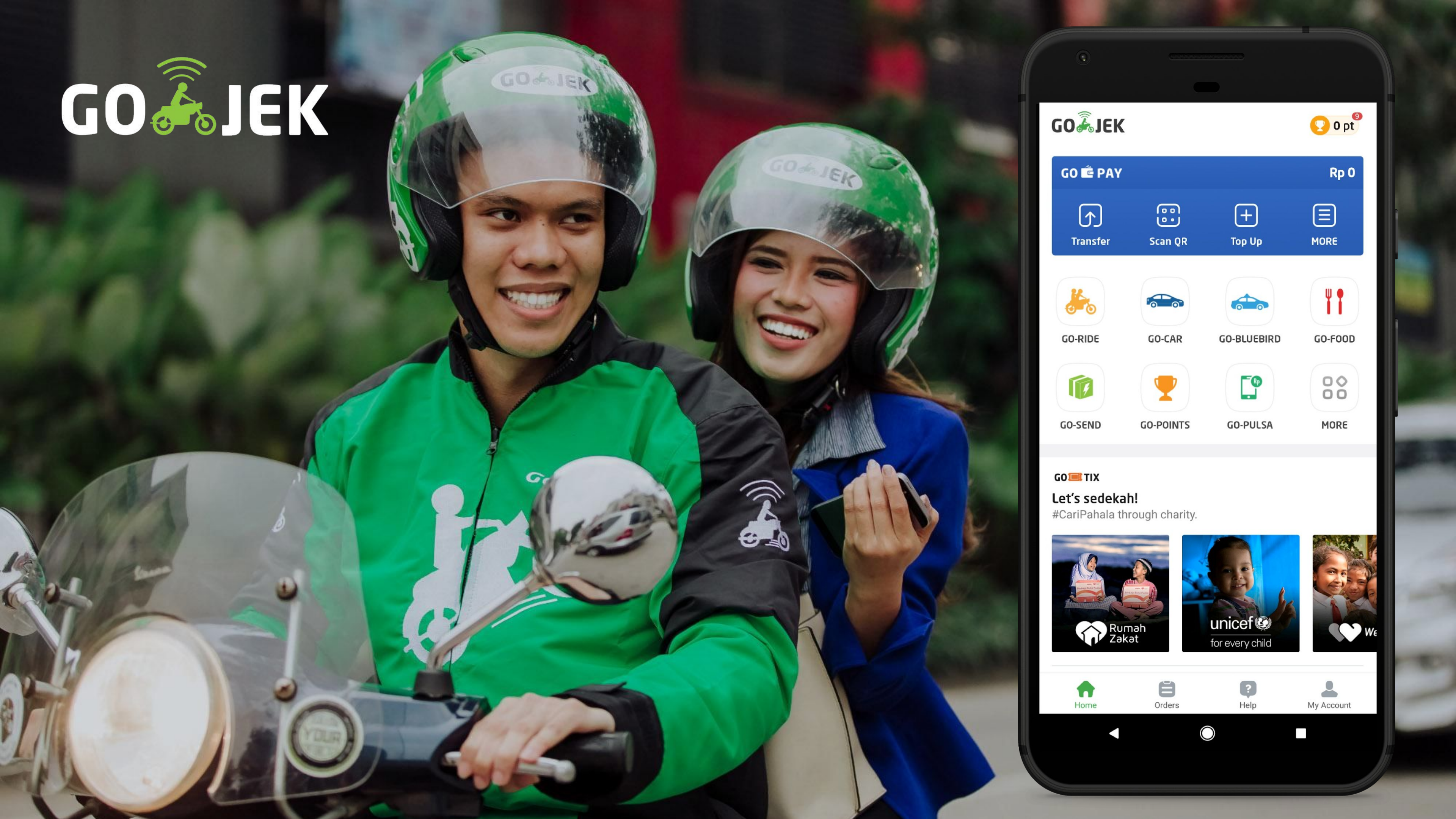
Takeaway Message

1. Kubernetes + Kubeflow is a really good platform for ML
2. Feast (Feature Store) + Kubeflow lets data scientists rapidly iterate on models



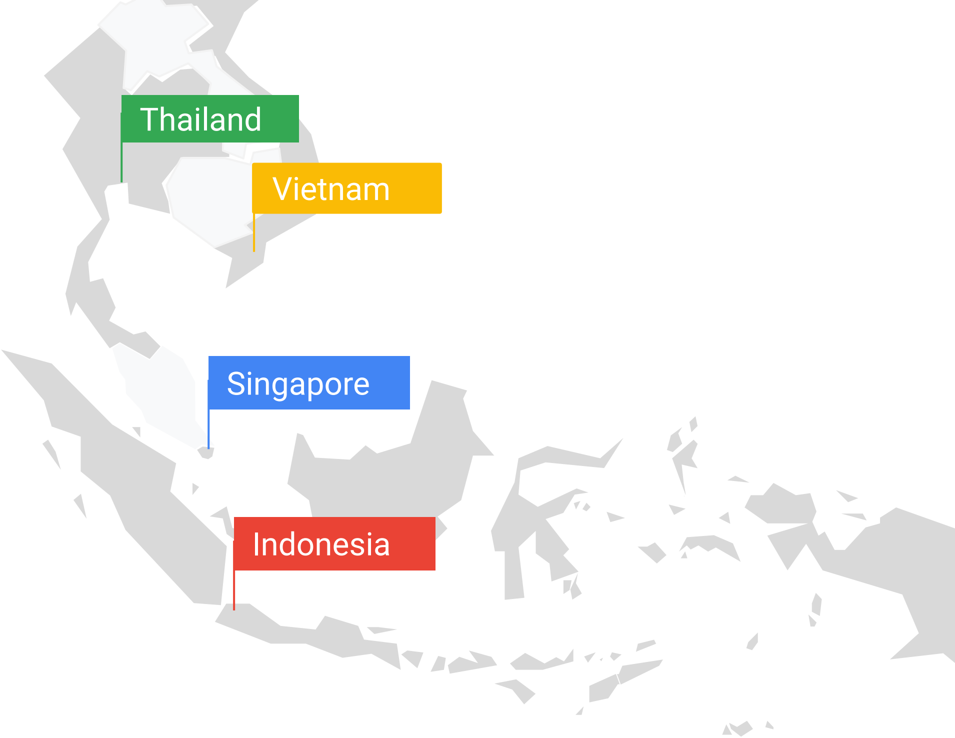
Agenda

- How K8s and Kubeflow empower companies like GOJEK to build ML platforms
- Demo - Using Feast + Kubeflow to build & deploy from notebooks
- What is Kubeflow


GO-JEK
GO-JEK
0 pt
GO PAY
Rp 0
Transfer
Scan QR
Top Up
MORE
GO-RIDE
GO-CAR
GO-BLUEBIRD
GO-FOOD
GO-SEND
GO-POINTS
GO-PULSA
MORE
GO TIX
Let's sedekah!
#CariPahala through charity.
Rumah Zakat
unicef
for every child
Home
Orders
Help
My Account

Our scale

Operating in 4 countries throughout Southeast Asia



125m+ app downloads

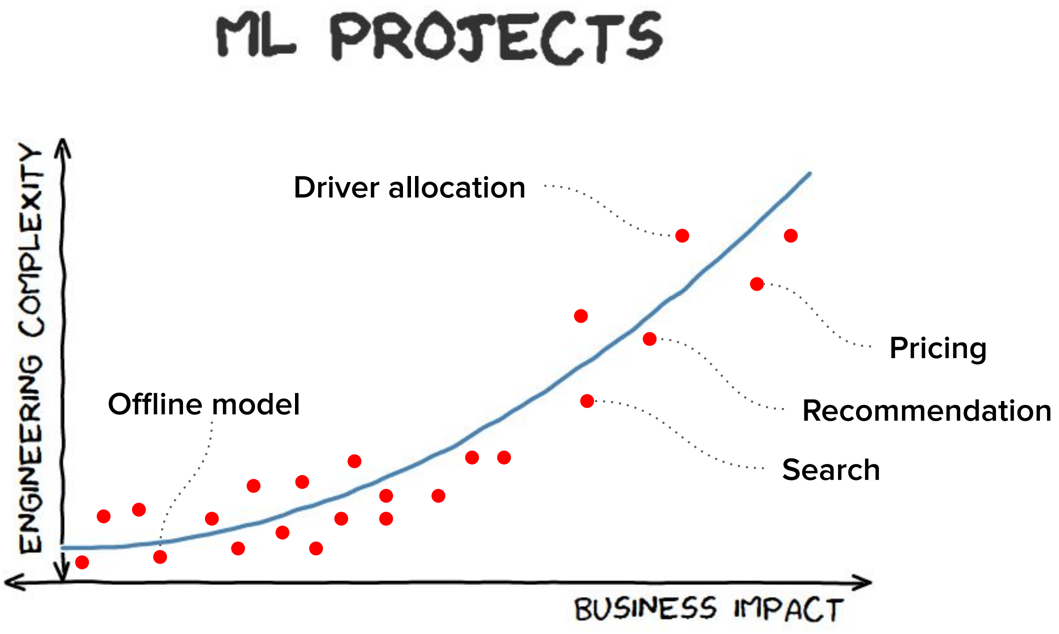
+400k merchants

4 countries

2m+ drivers

100m+ monthly bookings

Our Data



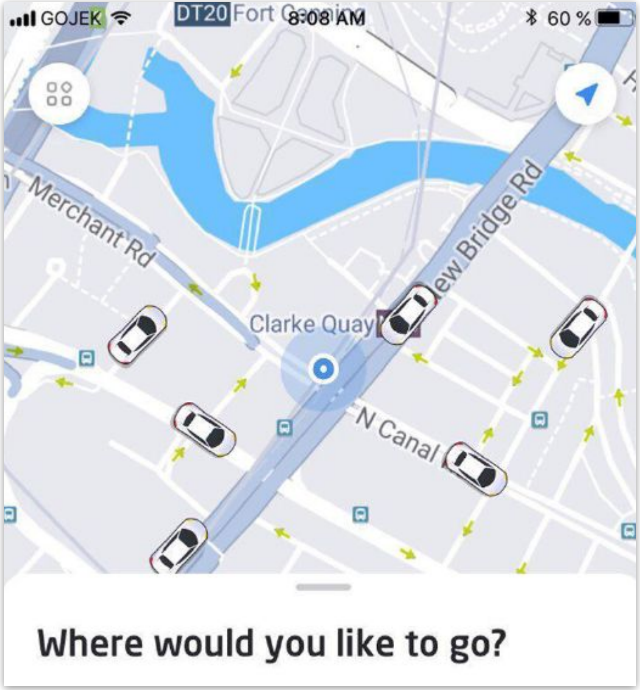
Data science requirements

- High abstraction
- Rapid iteration and experimentation
- Customizable workflows

Engineering requirements

- Integrates with existing systems (requires escape hatches)
- Able to operate at scale
- Easy to maintain and debug
- Easy to extend and build on

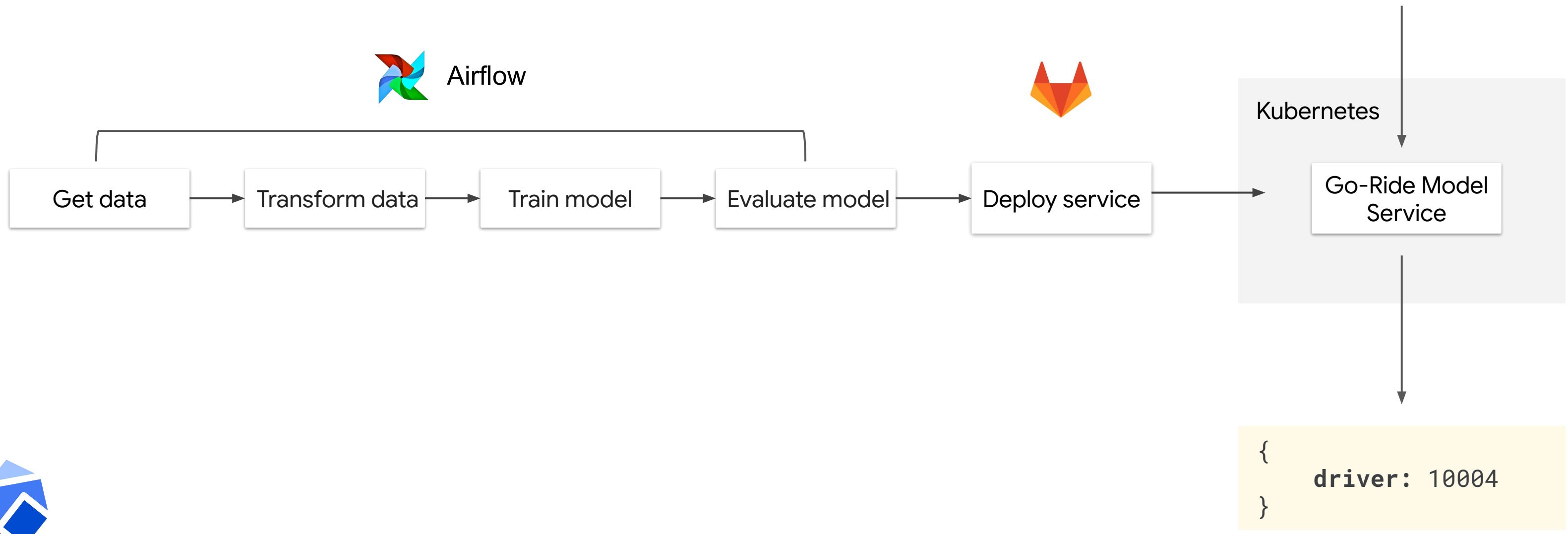
Finding the right driver



Let’s use Kubernetes

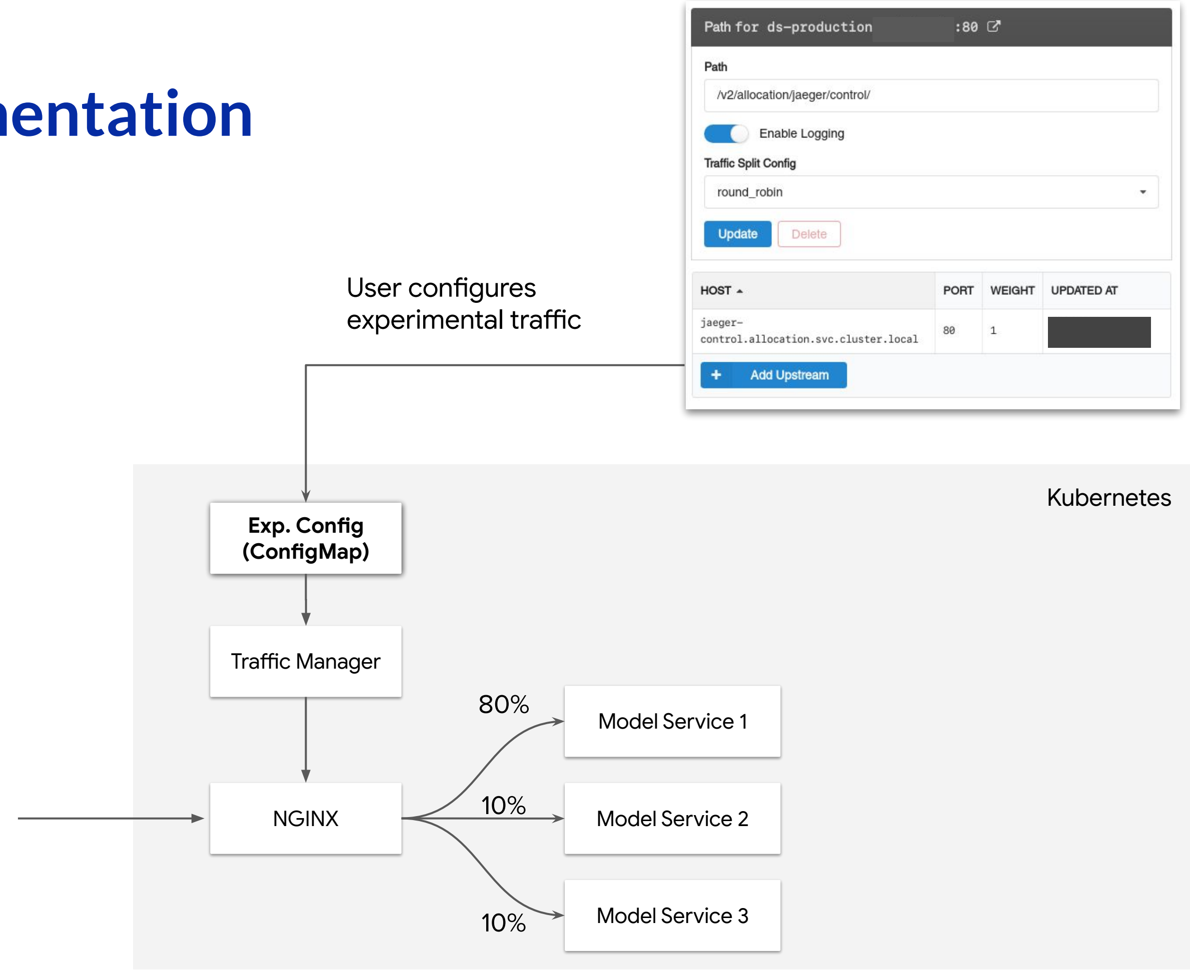
- Dependency management
- Container orchestration
- Process isolation


```
{
    customer: 12345,
    drivers: [
        10001,
        10002,
        10003,
        10004,
        10005
    ]
}
```

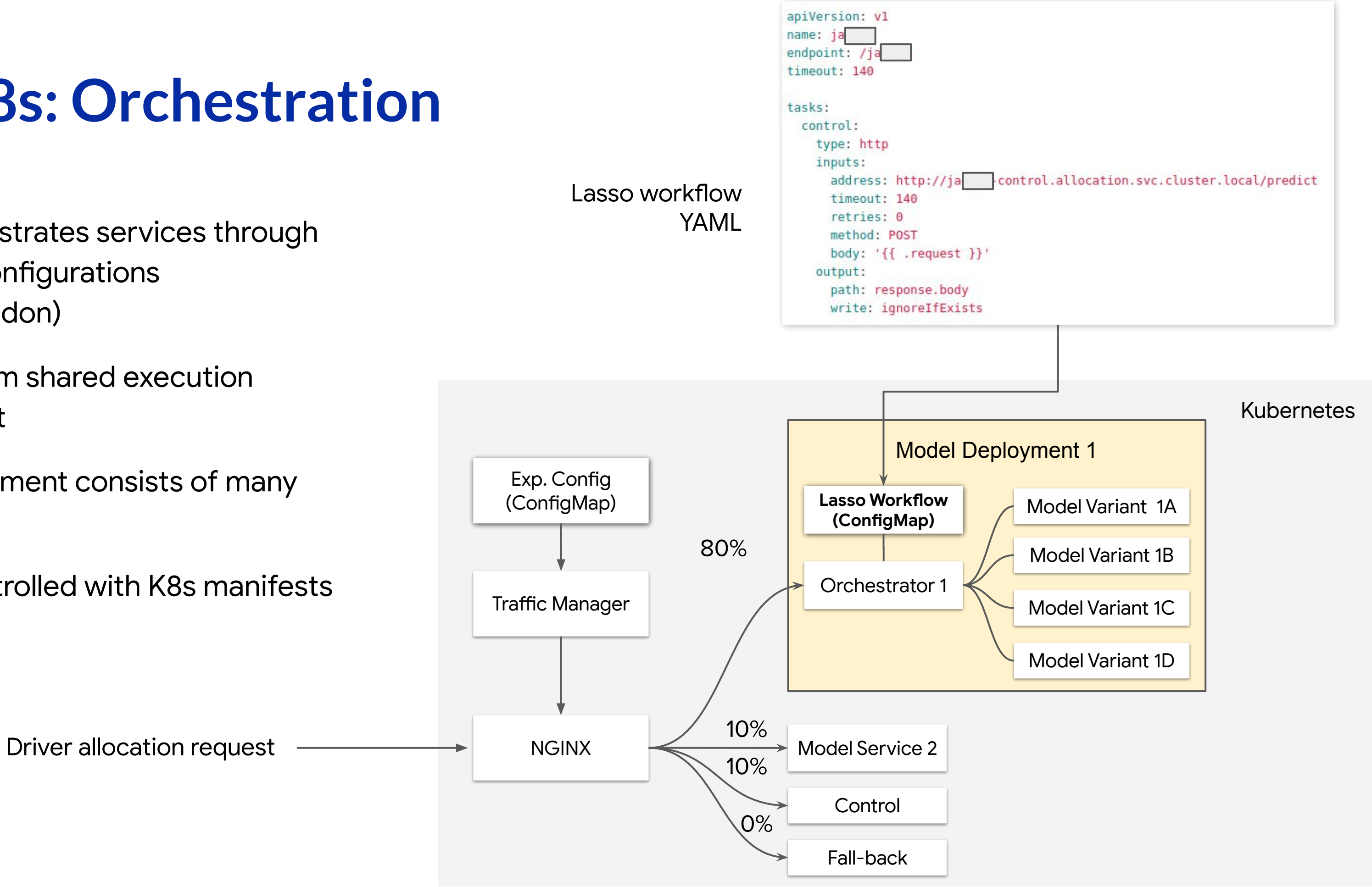
ML on K8s: Experimentation

- Experimentation manager required interface with ingress
- Kubernetes control plane provides common API
- Easy to develop to
- Encourages loose coupling between services



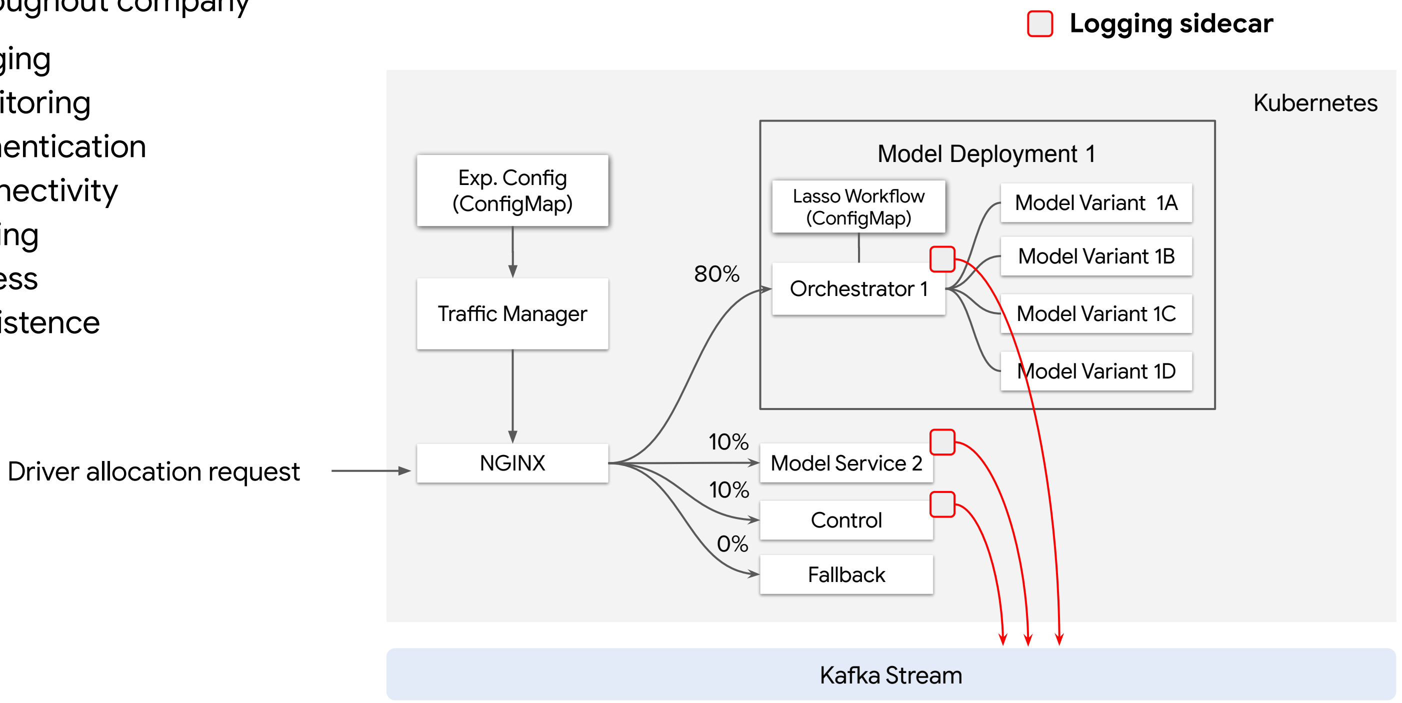
ML on K8s: Orchestration

- Lasso orchestrates services through workflow configurations (see also Seldon)
- Benefits from shared execution environment
- Each deployment consists of many models
- Version controlled with K8s manifests



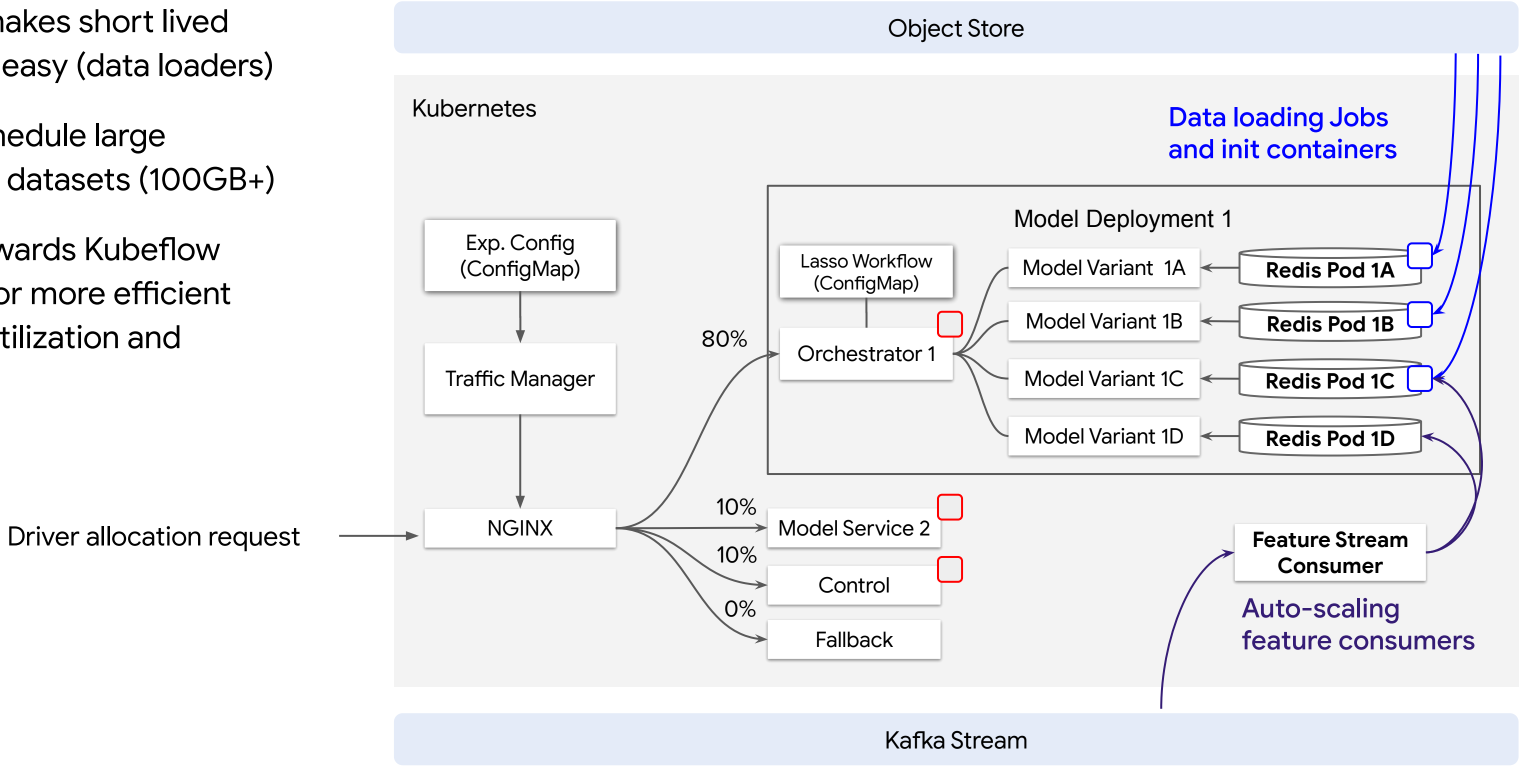
ML on K8s: Economies of scale

- Leverage components and tooling throughout company
  - Logging
  - Monitoring
  - Authentication
  - Connectivity
  - Tracing
  - Access
  - Persistence

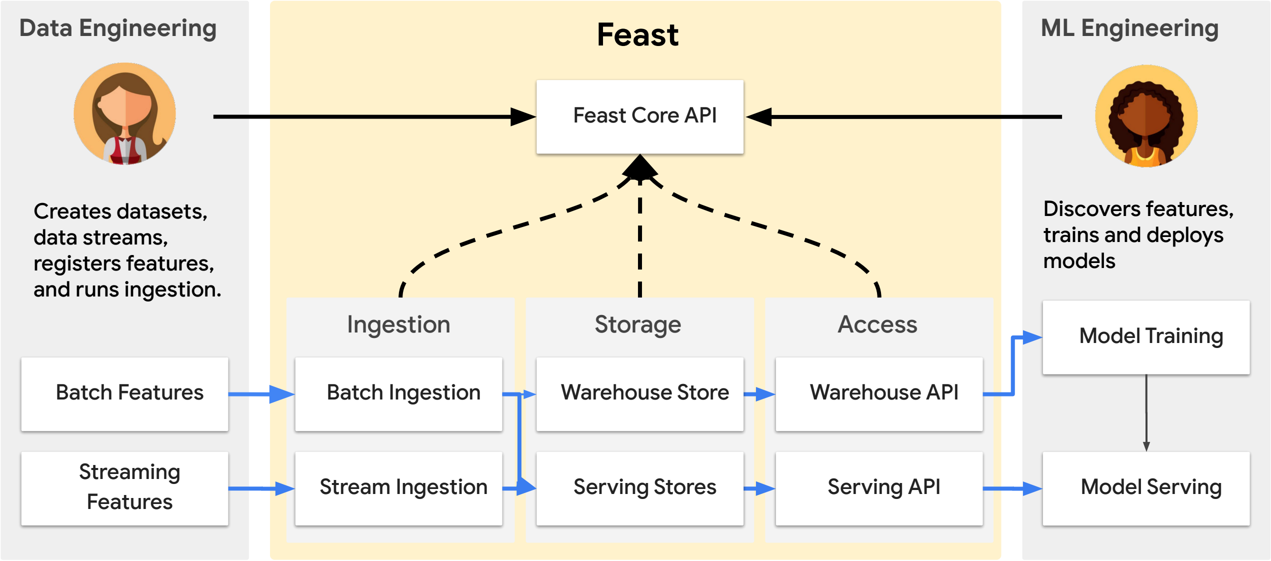


ML on K8s: Workloads

- Jobs API makes short lived processes easy (data loaders)
- Able to schedule large immutable datasets (100GB+)
- Moving towards Kubeflow Pipelines for more efficient resource utilization and tracking



ML on K8s: Feature Store (Feast)

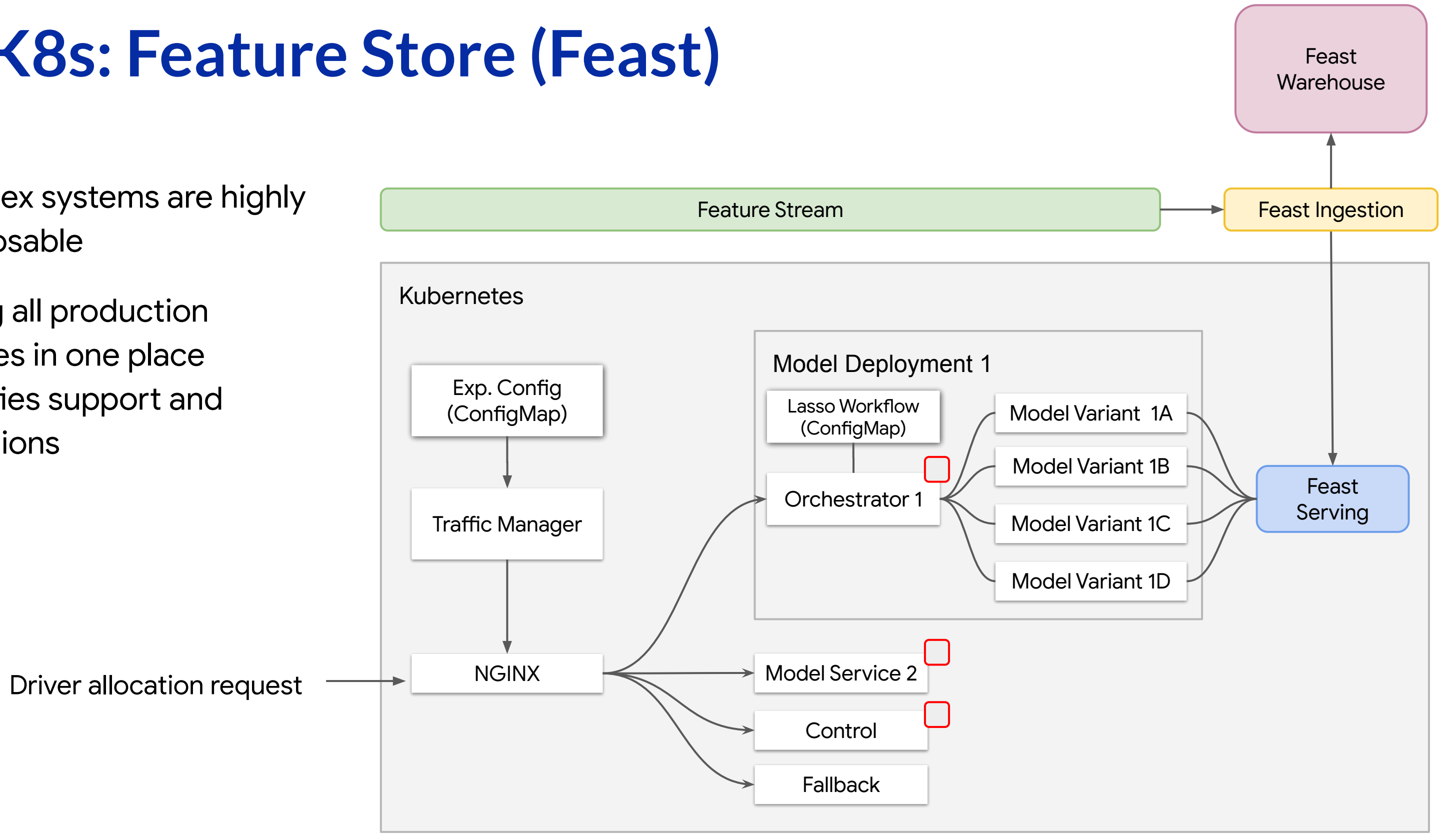


- Encourages feature discovery and reuse through centralization
- Prevents training-serving skew
- Provide scalable storage of feature data for serving and training

github.com/gojek/feast/

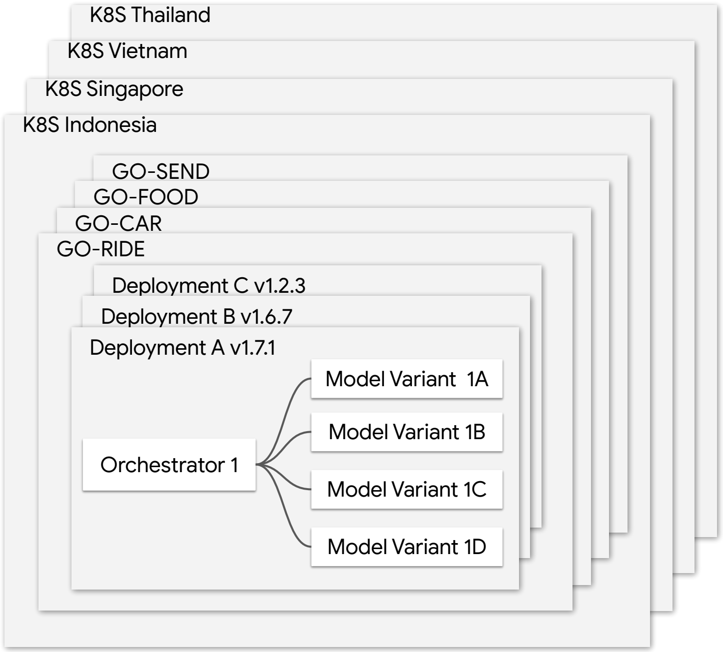
ML on K8s: Feature Store (Feast)

- Complex systems are highly composable
- Having all production services in one place simplifies support and operations



ML on K8s: Rapid expansion

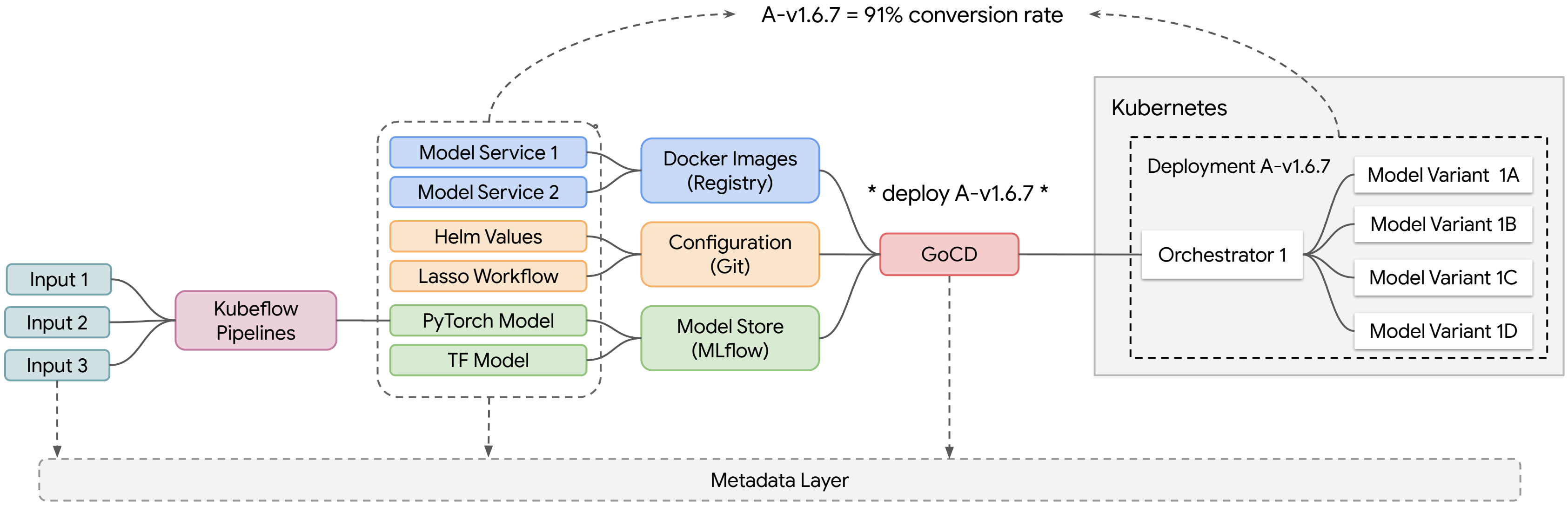
- Scaling to new markets required very large service-to-engineer ratio
- GitOps approach allowed us to increase our leverage and expand to new markets
- Terraform for all infrastructure (even Helm deployments)
- CD for model deployments



http://bit.ly/gojek-olympus (by Ravi Suhag)

ML on K8s + Kubeflow: Traceability

- Generalize & parameterize complete ML life cycle
- Track artifact combination as deployment version
- Measure version against experiment



ML on K8s: The good parts

- Large ecosystem of ML frameworks built on top of Kubernetes
- Consistent API simplifies developing complex systems
- Workloads benefit from intelligent scheduling, resource utilization, and dependency management.
- Single production environment simplifies operations
- GitOps allows for high leverage and portability
- Artifact based versioning and tracking allows for traceable experiments

ML on K8s: The rough parts

- Multi tenancy
- Stateful systems aren’t there yet
- Leaky abstractions (CRDs / annotations exposed to users)

ML on K8s: What's next

- Simplify the end-to-end user experience
- Metadata tracking (Kubeflow)
- Istio integration

Kelsey Hightower
@kelseyhightower

Kubernetes is a platform for building platforms. It's a better place to start; not the endgame.

10:04 PM · Nov 27, 2017

237 Retweets **676** Likes

Kubeflow is an open, Kubernetes native platform for ML

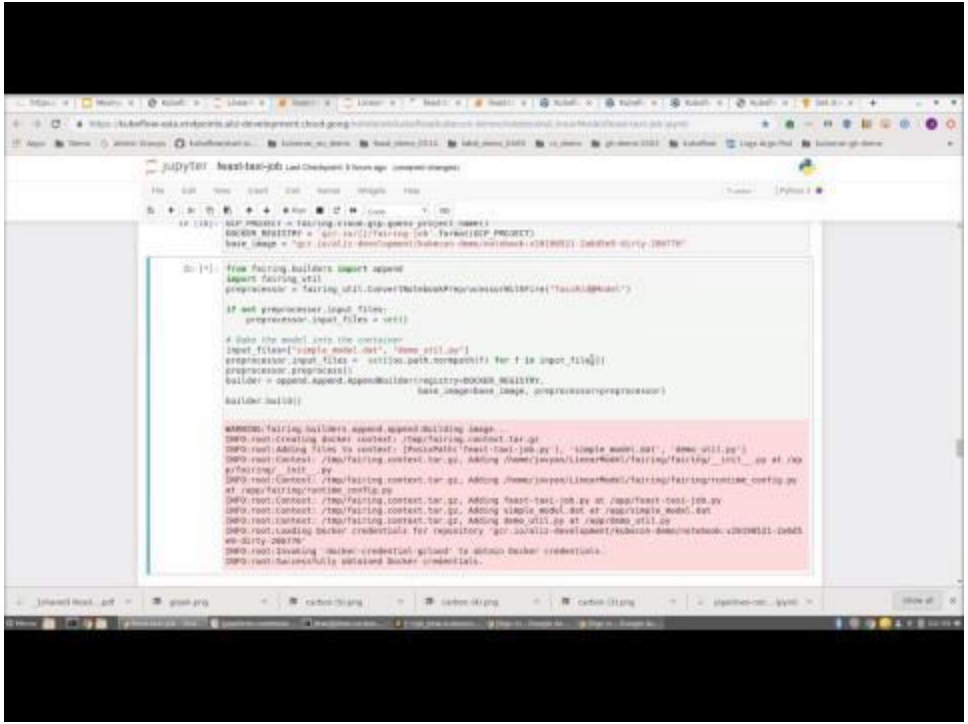
Make it Easy for Everyone to Develop, Deploy and Manage Portable, Distributed ML on Kubernetes

Demo setup

Using Feast and Kubeflow to build and deploy models

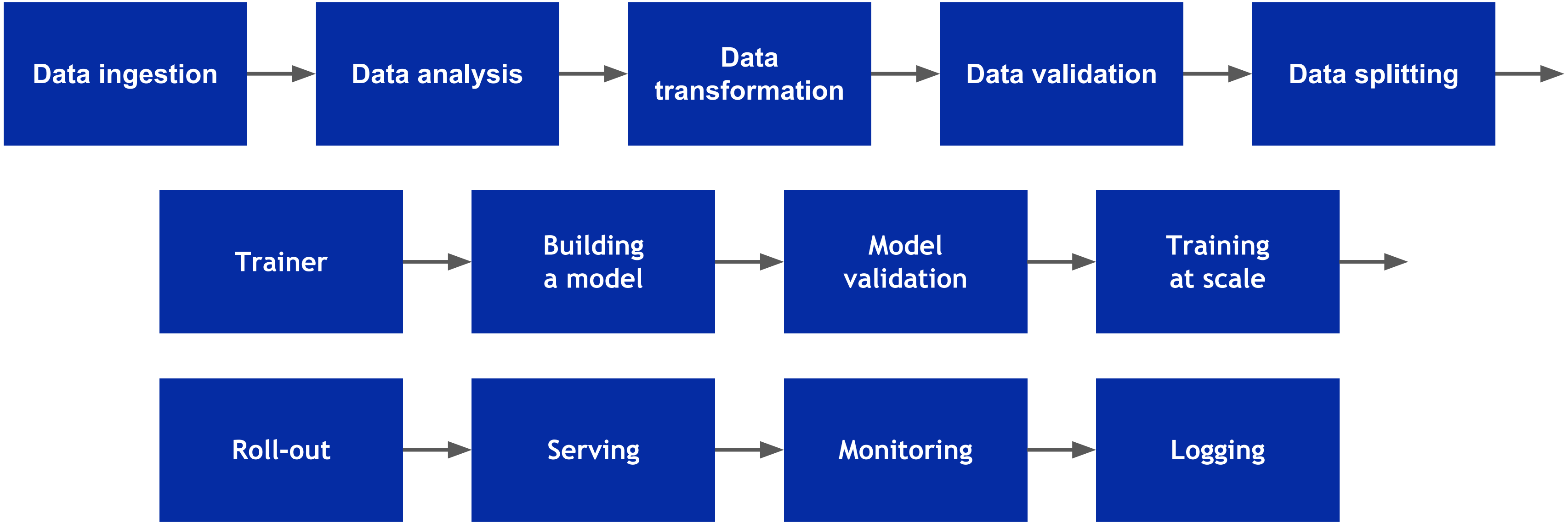
- Fetch data from feast for training
- Develop models in a notebook
- Deploy models on K8s
- Fetch data at serving time for inference

Video of Demo



Kubeflow

ML Development Workflow

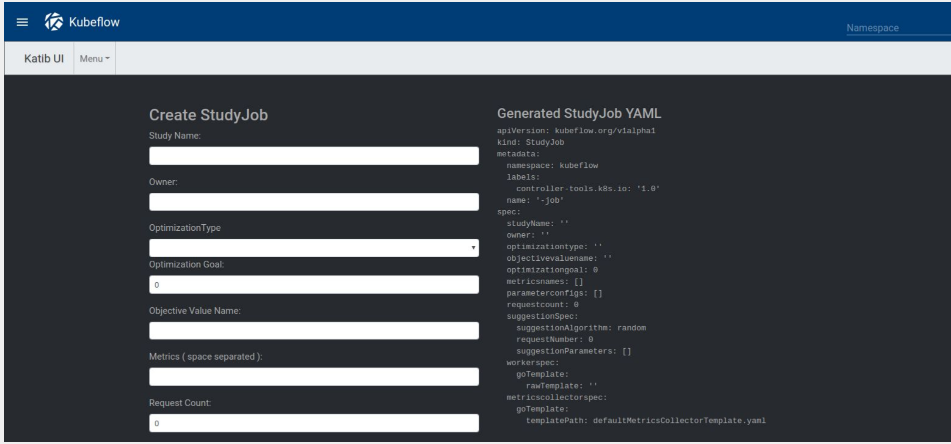


Kubeflow makes it easy to run these steps on Kubernetes

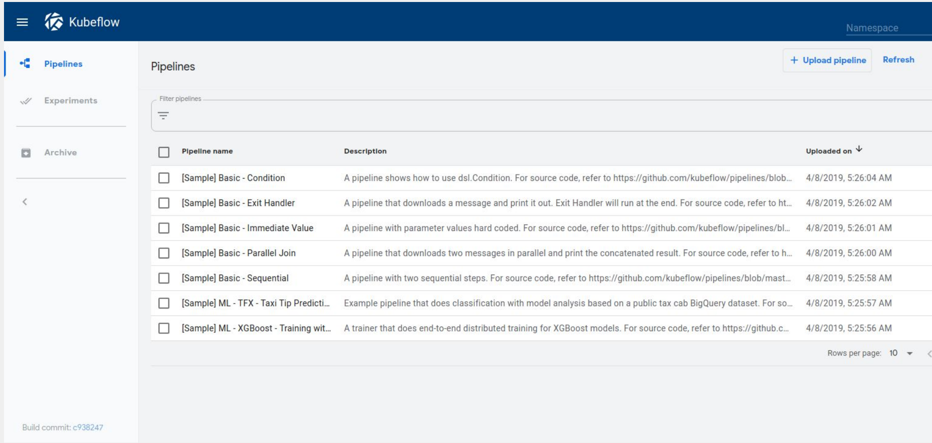
Core Tenets

- Kubeflow makes it easy to run ML applications on Kubernetes
  - e.g. notebooks, HP tuning, pipelines, model servers, etc...
- **Composable -** Use the libraries/frameworks of your choice
- **Scalable -** number of users & workload size
- **Portable -** on prem, public cloud, local

Hyperparameter Tuning



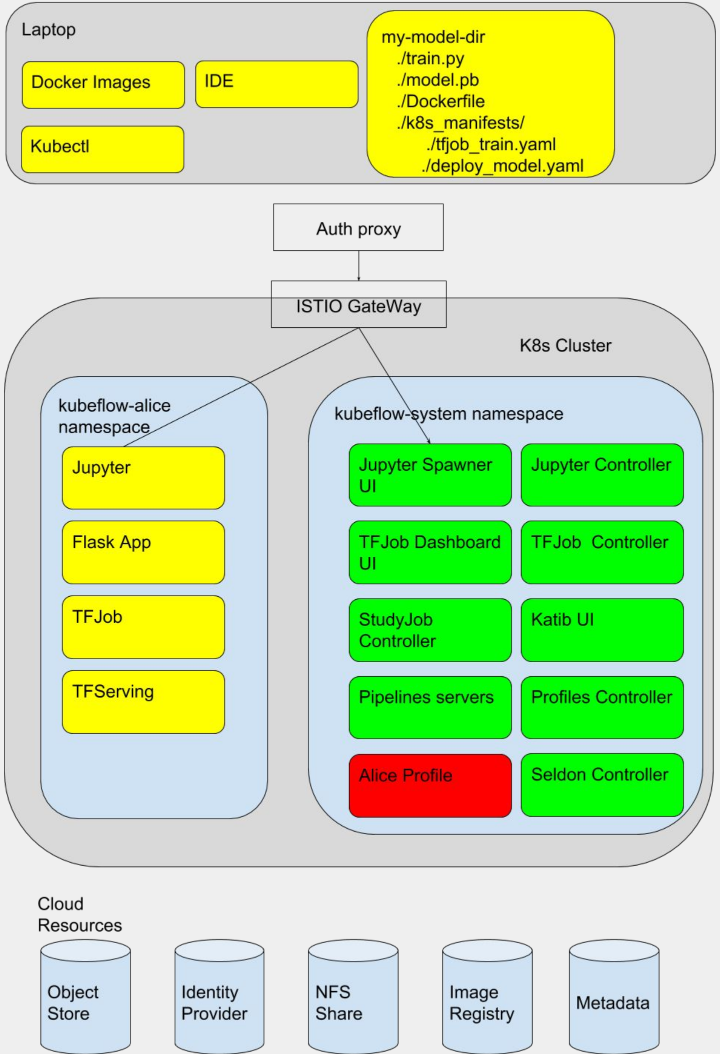
Pipelines



Kubeflow Architecture

System Diagram

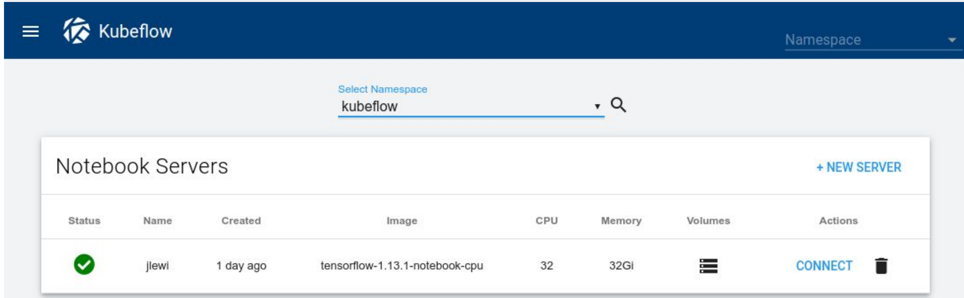
- Kubeflow is a collection of microservices
  - Will use ISTIO for service mesh in 0.6
- Users/teams consume Kubeflow in their own namespace

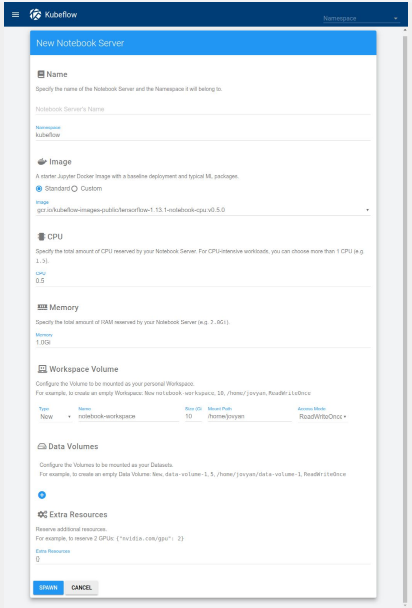


OSS Momentum!

Goal: low bar; high ceiling

- Day 0 focus on model development
  - Use UIs to launch notebooks
  - Python SDK (fairing) for training / deploying models
- Day 0 start with the infrastructure (Kubernetes, ISTIO, etc...) that you can ride into production
- Day N leverage K8s to scale
  - Use the same infrastructure as non-ML applications
  - Build a single infrastructure team





Call To Action

- Install Kubeflow - https://www.kubeflow.org/docs/started/
- Install Feast - https://github.com/gojek/feast/blob/master/docs/install.md
- Try Fairing + Feast:
    - https://github.com/gabrielwen/LinearModel

Kubeflow Talks (bit.ly/kf_calendar)

- **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)
- **Toward Kubeflow 1.0, Bringing a Cloud Native Platform for ML to Kubernetes** - *Wednesday May 22 11:55 - 12:30*; David Aronchick, Microsoft & Jeremy Lewi Google (https://sched.co/MPax)
- **Building Cross-Cloud ML Pipelines with Kubeflow with Spark & TensorFlow** - *Wednesday May 22 14:00 - 14:35*; Holden Karau, Google & Trevor Grant, IBM (https://sched.co/MPaZ)
- **Managing Machine Learning Pipelines In Production with Kubeflow with Devops** - *Wednesday May 22 14:40-14:35* - David Aronchick, Microsoft (https://sched.co/MPaZ)
- **Large Scale Distributed Deep Learning with Kubernetes Operators -** *Wed May 22 15:55 - 16:30;* Yuan Tang, Ant Financial & Yong Tang MobileIron (https://sched.co/MPaT)
- **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)

Kubeflow

Thank You

www.kubeflow.org

github.com/gojek/feast

github.com/gabrielwen/LinearModel