PORTABLE MACHINE LEARNING DEPLOYMENTS IN 

Salman Iqbal
SALMAN IQBAL

@soulmani iqbal
SOFTWARE DEVELOPER
SOFTWARE DEVELOPER

❤️ BUILDING ON ☁
SOFTWARE DEVELOPER

❤️ BUILDING THINGS ON ☁

DATA SCIENTIST
SOFTWARE DEVELOPER

❤️ BUILDING THINGS ON ☁

❌ DATA SCIENTIST
UK MEAN TEMPERATURE FOR SEPTEMBER 2019?
~ Met Office
13.1 °C
UK Maximum Temperature for September 2019?

~ Met Office
27.7 °C

Weybourne (Norfolk)
AGENDA
THANKS DANIELE!

@danielepolencic
THANKS ECMWF!
13 layer policy network
13 layer policy network
30 million positions
13 layer policy network
30 million positions
1,920 CPUs & 280 GPUs
CHALLENGES
1. Compile once run anywhere
1. Compile once run anywhere
1. Works on my machine
- Big Data
- Scale with team and model
- Cost for money
- Quick fix
BIG DATA

SCALE WITH TEAM AND MODEL

COST FOR MONEY

QUICK FIX
BIG DATA

SCALE WITH TEAM AND MODEL

COST FOR MONEY

QUICK FIX
BIG DATA
SCALE WITH TEAM AND MODEL
COST FOR MONEY
QUICK FIX
BIG DATA
SCALE WITH TEAM AND MODEL
COST FOR MONEY
QUICK FIX
2. COMPLEXITY
Training Data
Training Data

Test Data
Training Data

Test Data

Model
Training Data

Test Data

Model

Evaluate
Data collection

Training Data

Test Data

Model

Evaluate
Data collection

Feature extraction

Configuration

Data verification

Training Data
Test Data
Model
Evaluate
Training Data
Test Data
Model
Evaluate
Data collection
Feature extraction
Data verification
Machine resource management
Serving Infrastructure
Configuration
Analysis tools
Monitoring
LEARN DEVELOPMENT
LEARN OPS & INFRA
LEARN DATA SCIENCE

@soulmaniqbal
LEARN DEVELOPMENT

LEARN OPS & INFRA

LEARN DATA SCIENCE

@soulmaniqbal
LEARN DEVELOPMENT
LEARN OPS & INFRA
LEARN DATA SCIENCE

@soulmaniqbal
LEARN DEVELOPMENT
LEARN OPS & INFRA
LEARN DATA SCIENCE
@soulmaniqbal
DATA SCIENTIST — EXPECTATION
3. **TRAINING HUNDREDS OF MODELS**
Recommendation  Antispam  Churn rate
Recommendation v1
Recommendation v2
<table>
<thead>
<tr>
<th></th>
<th>Monday</th>
<th></th>
<th>Tuesday</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recommendation v1</td>
<td></td>
<td>Recommendation v1</td>
</tr>
<tr>
<td></td>
<td>Antispam v1</td>
<td></td>
<td>Antispam v1</td>
</tr>
<tr>
<td></td>
<td>Churn rate v1</td>
<td></td>
<td>Churn rate v1</td>
</tr>
<tr>
<td></td>
<td>Recommendation v2</td>
<td></td>
<td>Recommendation v2</td>
</tr>
<tr>
<td></td>
<td>Antispam v2</td>
<td></td>
<td>Antispam v2</td>
</tr>
<tr>
<td></td>
<td>Churn rate v2</td>
<td></td>
<td>Churn rate v2</td>
</tr>
<tr>
<td></td>
<td>Recommendation v3</td>
<td></td>
<td>Recommendation v3</td>
</tr>
<tr>
<td></td>
<td>Antispam v3</td>
<td></td>
<td>Antispam v3</td>
</tr>
<tr>
<td></td>
<td>Churn rate v3</td>
<td></td>
<td>Churn rate v3</td>
</tr>
<tr>
<td></td>
<td>Recommendation v4</td>
<td></td>
<td>Recommendation v4</td>
</tr>
<tr>
<td></td>
<td>Antispam v4</td>
<td></td>
<td>Antispam v4</td>
</tr>
<tr>
<td></td>
<td>Churn rate v1</td>
<td></td>
<td>Churn rate v1</td>
</tr>
</tbody>
</table>

**Days:**
- **Monday**
- **Tuesday**
Manual Automation 😂 😂
SLOW!
SLOW!

ERROR PRONE!!
SLOW!
ERROR PRONE!!
BORING!!!
WHAT CAN YOU DO?
WHAT IF...
YOU COULD USE ☁️?
RENT KIT BY THE HOUR
RENT KIT BY THE HOUR

BIGGER MACHINES
RENT KIT BY THE HOUR
BIGGER MACHINES
SCALE WITH THE TEAM
WHAT IF...
RENT KIT BY THE HOUR
RENT KIT BY THE HOUR
RENT FLEET OF MACHINE BY THE HOUR
RENT FLEET OF MACHINE BY THE HOUR

BIGGER MACHINES
Rent fleet of machine by the hour

Bigger machines

Mixed nodes w/ CPUs, GPUs, TPUs
Rent fleet of machine by the hour with mixed nodes w/ CPUs, GPUs, TPUs. Scale with the team.
Rent fleet of machine by the hour.

Mixed nodes w/ CPUs, GPUs, TPUs.

Scale with the team.
Rent fleet of machine by the hour
mixed nodes w/ CPUs, GPUs, TPUs
scale with the company
WHAT IF ...
THIS IS READY TO BE USED?
WHAT IS IT?
also
STILL SINGLE JOB...
HOW CAN I LOVE JOBS?
Kubernetes
Kubernetes

Kubeflow

Kubernetes
Kubernetes

Kubeflow

PyTorch

Kubernetes
ARCHITECTURE
MASTER
MASTER

PARAMETER SERVER

PARAMETER SERVER
MASTER

PARAMETER SERVER

PARAMETER SERVER

WORKER

WORKER

WORKER

WORKER

WORKER
MEANWHILE
DOES K8S RUN PYTHON?
NOPE!
CONTAINER ORCHESTRATOR
CONTAINER
python + JSON

PY

{}
Python + JSON + TensorFlow
LANGUAGE AGNOSTIC
DEMO...
MINIKF
by AARIKTO
Chicago Taxi
Chicago Taxi
100M Trips
Chicago Taxi
100M Trips
fare, tripstartmonth, tripstarthour
Chicago Taxi
100M Trips
fare, tripstartmonth, tripstarthour
Out: tips > 20%
Pipeline
1. data validation
1. Data Validation
2. Data Preprocessing
1. DATA VALIDATION
2. DATA PREPROCESSING
3. MODEL TRAINING
1. Data Validation
2. Data Preprocessing
3. Model Training
4. Model Analysis
1. Data Validation
2. Data Preprocessing
3. Model Training
4. Model Analysis
5. Prediction
Step 1
1. Clone disk from snapshot
2. Do initial analysis
3. Snapshot

Step 2
4. Clone disk of Step 1
5. Transform data
6. Snapshot

Step 3
7. Clone disk of Step 2
8. Train model
9. Snapshot

Arrikto

Object Store
MORE...
NETWORK TOPOLOGY
K8s General Purpose Software
TF JOB
apiVersion: kubeflow.org/v1
kind: TFJob
metadata:
  generateName: tfjob
  namespace: kubeflow
spec:
tfReplicaSpecs:
  PS:
    replicas: 1
    restartPolicy: OnFailure
  .
  .
Worker:
    replicas: 3
    restartPolicy: OnFailure
    template:
      spec:
        containers:
          - name: tensorflow
            image: gcr.io/your-project/your-image
            command:
              - python
              - -m
              - trainer.task
              - --batch_size=32
              - --training_steps=1000
Fairing Attribute
MACHINE LEARNING AT SCALE WITH K8S
Bloomberg
6000 instances of Apache Solr ~ 1000 servers
TRAINING ML WITH 2500 NODES
HORIZONTAL POD AUTOSCALER
CLUSTER AUTOSCALER
256 P100
128000 PREEMPTIBLE VCPUS
256 P100 $400/HR
128000 PREEMPTIBLE VCPUS $1280/HR
256 P100 $400/HR
128000 PREEMPTIBLE VCPUS $1280/HR
320,000 CORES
320,000 CORES

4,300 PROJECTS
320,000 cores
4,300 projects
330 petabytes
3,300 users
3,300 users
10,000 hypervisors
3,300 users
10,000 hypervisors
210 Kubernetes clusters
NOT FREE
Docker: ~950K loc
Kubernetes: ~3.1M
Kubeflow: ~830K
Pipelines: ~220K
KFServing: ~3.1M~
Kubernetes
Containers
Kubeflow - run anywhere Demo
WHAT'S NEXT?
Resources

- https://www.kubeflow.org/
- https://kubernetes.io/
- https://learnk8s.io/academy/
- https://deploy.kubeflow.cloud
- bit.ly/kubeflow-taxi
- https://www.arrikto.com/
Resources:

- https://kubernetes.io/case-studies/cern/
- Kubernetes @ CERN
- Kubernetes @ CERN presentation
THANKS!

- @soulmaniqbal
- @learnk8s