BBC
Design + Engineering

DATA SCIENCE FRAMEWORKS AND MANAGED SERVICES: WHEN TO AVOID THE SHINY NEW TOYS

JON TUTCHER, BBC DATALAB, PYDATA LONDON

14 JULY 2019

Airflow DAGs Data Profiling Browse Admin Docs About

On DAG: pydata_london

schedule: None

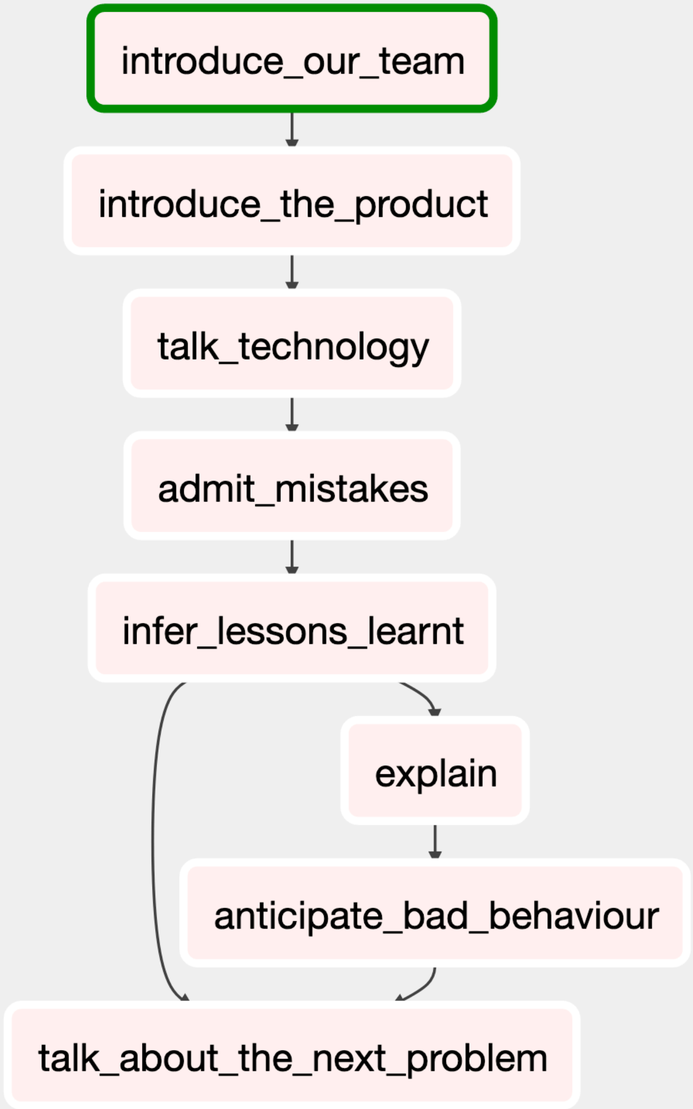
Graph View | Tree View | Task Duration | Task Tries | Landing Times | Gantt | Details | Code | Trigger DAG | Refresh | Delete

running Base date: 2019-07-13 18:38:04 Number of runs: 25 Run: manual__2019-07-13T18:38:03.302273+00:00 Layout: Top->Bottom Go

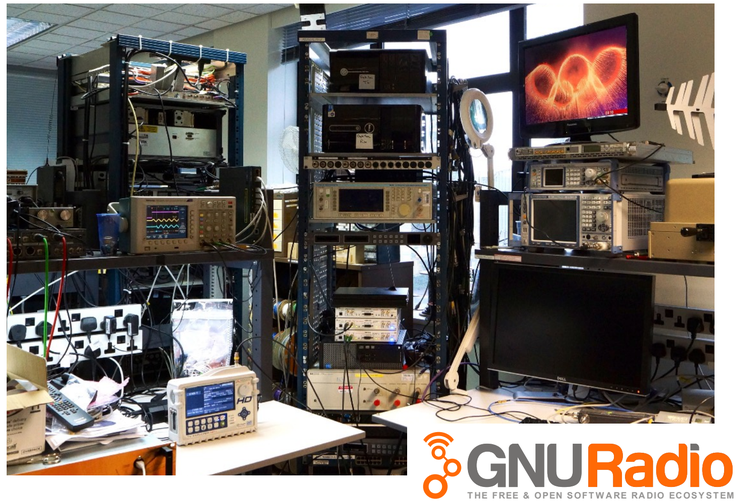
Search for...

PythonOperator

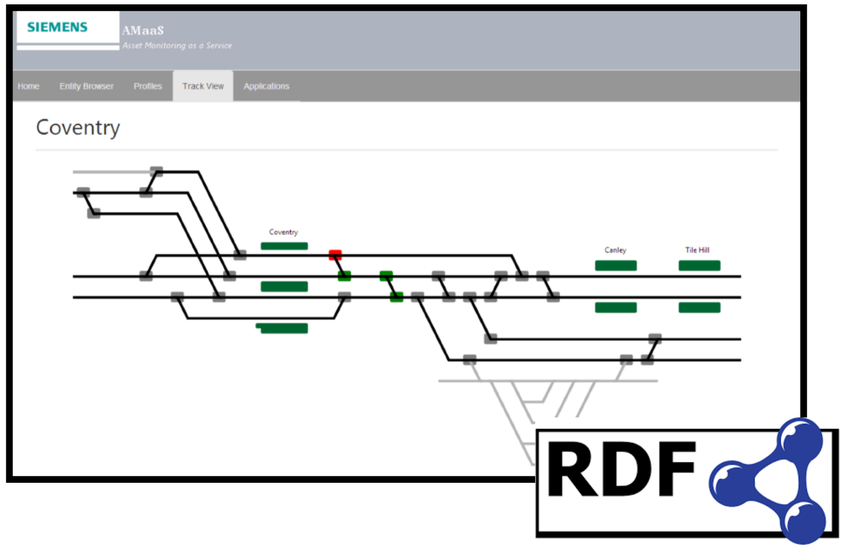
success running failed skipped up_for_reschedule up_for_retry queued no_status




HYDROGEN PIONEER
POWERED BY
LabVIEW


GNURadio
THE FREE & OPEN SOFTWARE RADIO ECOSYSTEM


PROD. NO. CAKE
OBJECT 18
SHOT 30
ROLL 1 A+B
DATE 07/04/16
SOUND
PROD. CO.
DIRECTOR JASMINE COX
CAMERA MAX LEONARD + MATT SHOTTON
node


SIEMENS
AMaaS
Asset Monitoring as a Service
Home
Entity Browser
Profiles
Track View
Applications
Coventry
Coventry
Canley
Tile Hill
RDF


1:00:00
1:00:20
1:00:40
1:01:00
1:01:20
1:01:40
Intro
Verse 1
Chorus
Verse 2
Chorus
Mid 4
Verse 3
The Roots
Mellow My Man
DGC/GEFFEN.

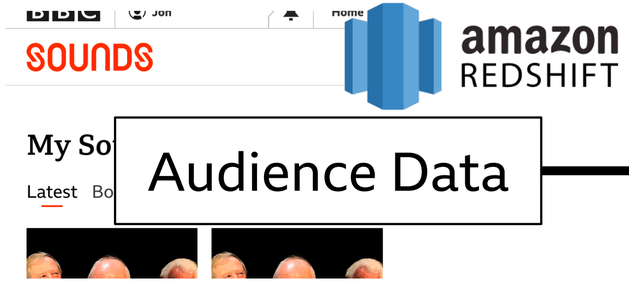
dataLab

BBC

Content

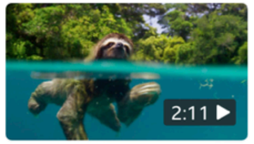

```
<synopses>
  <synopsis length="short">
    The award-winning actor Vanessa Redgrave on her new play. One woman's anoı
  </synopsis>
  <synopsis leng
    Award-winnin
  </synopsis>
  <synopsis leng
    Olivier, Tony and Academy Award winner Vanessa Redgrave on her new stage p

    Our listener, Sarah tells us about being in recovery from a lifetime of anorexia an
```


Content Metadata



Audience Data

BBC datalab



Swimming sloth

Nature & Environment

All the things people say when you're a young mum!

The Social

Things not to say to body builders

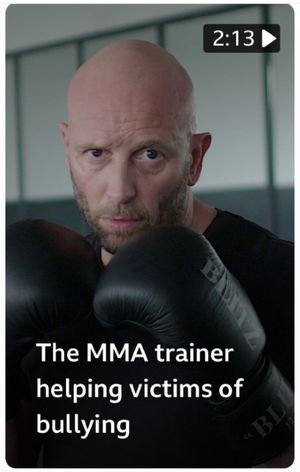
Beauty & Style

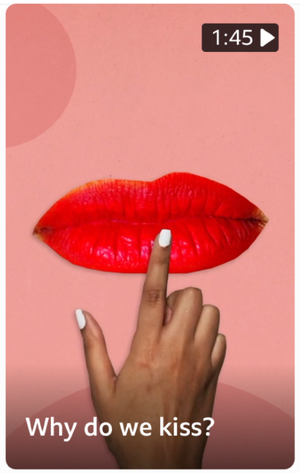
9:41

TUE 7 MAY

Good Morning

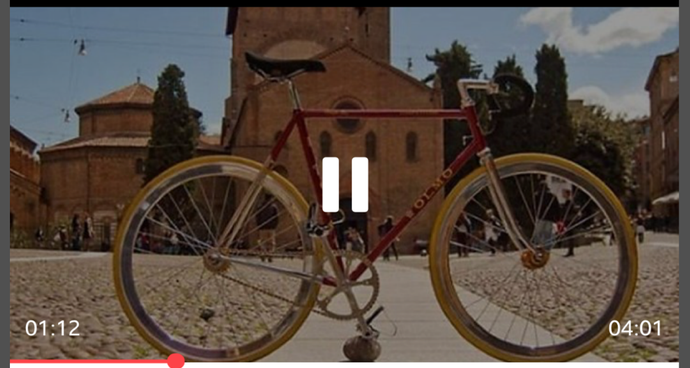
Videos of the day





Money

Watch | Discover | You



The city that gives you free beer for cycling

42 minutes ago

BBC World Hacks

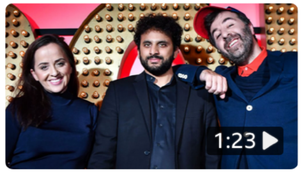
Related

How to eat carbs and stay healthy

Food & Drink

Five compelling reasons why we all need to sleep more

Health & Wellbeing

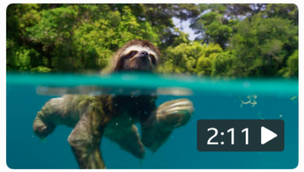


The malicious influence of the Spice Girls

Stand-up

9:41

Inspiring | Cancel



Swimming sloth

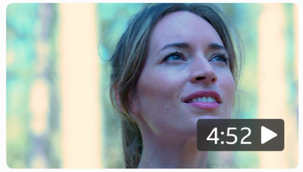
Nature & Environment

All the things people say when you're a young mum!

The Social

Things not to say to body builders

Beauty & Style



Tips for success by the youngest British woman to climb Everest

Top Tips

How to make extra crunchy roast potatoes

Food

How your phone can save your life

Ideas

Watch | Discover | You

BBC+

GETTING STARTED

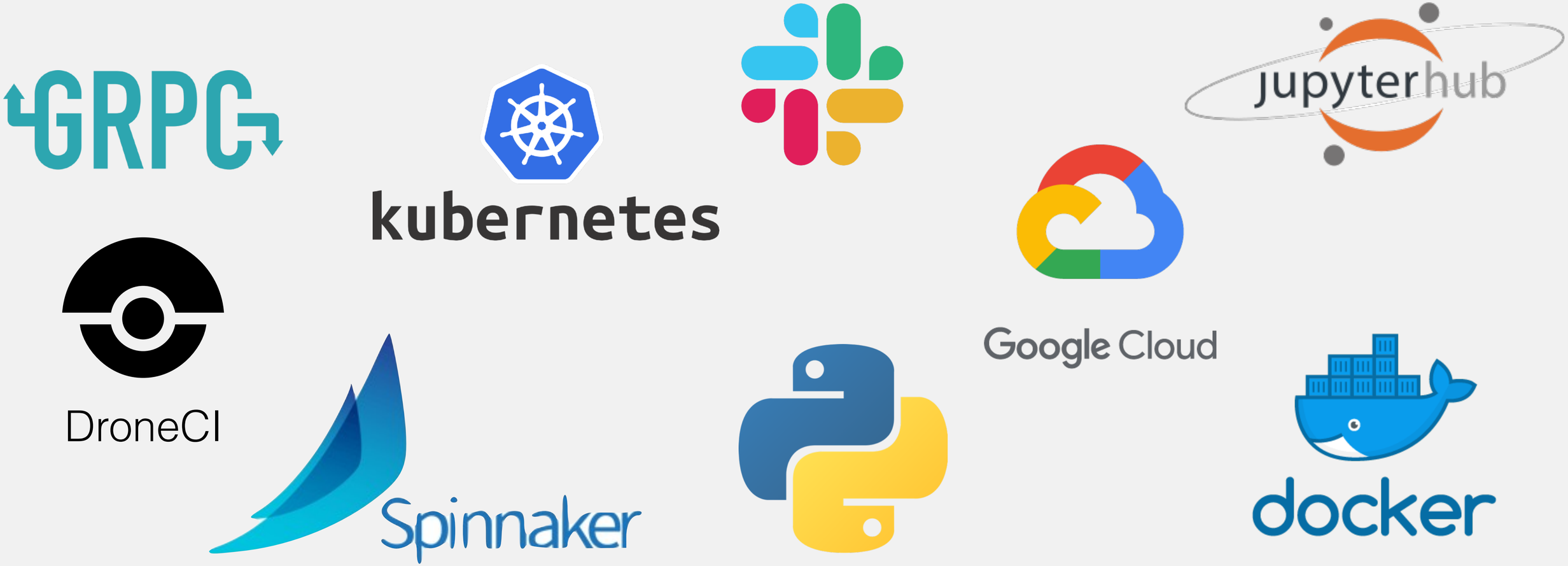
*"We'll have other customers - we need to build a **platform**"*

"We're a new, independent team - the bosses want us to try some new technology"

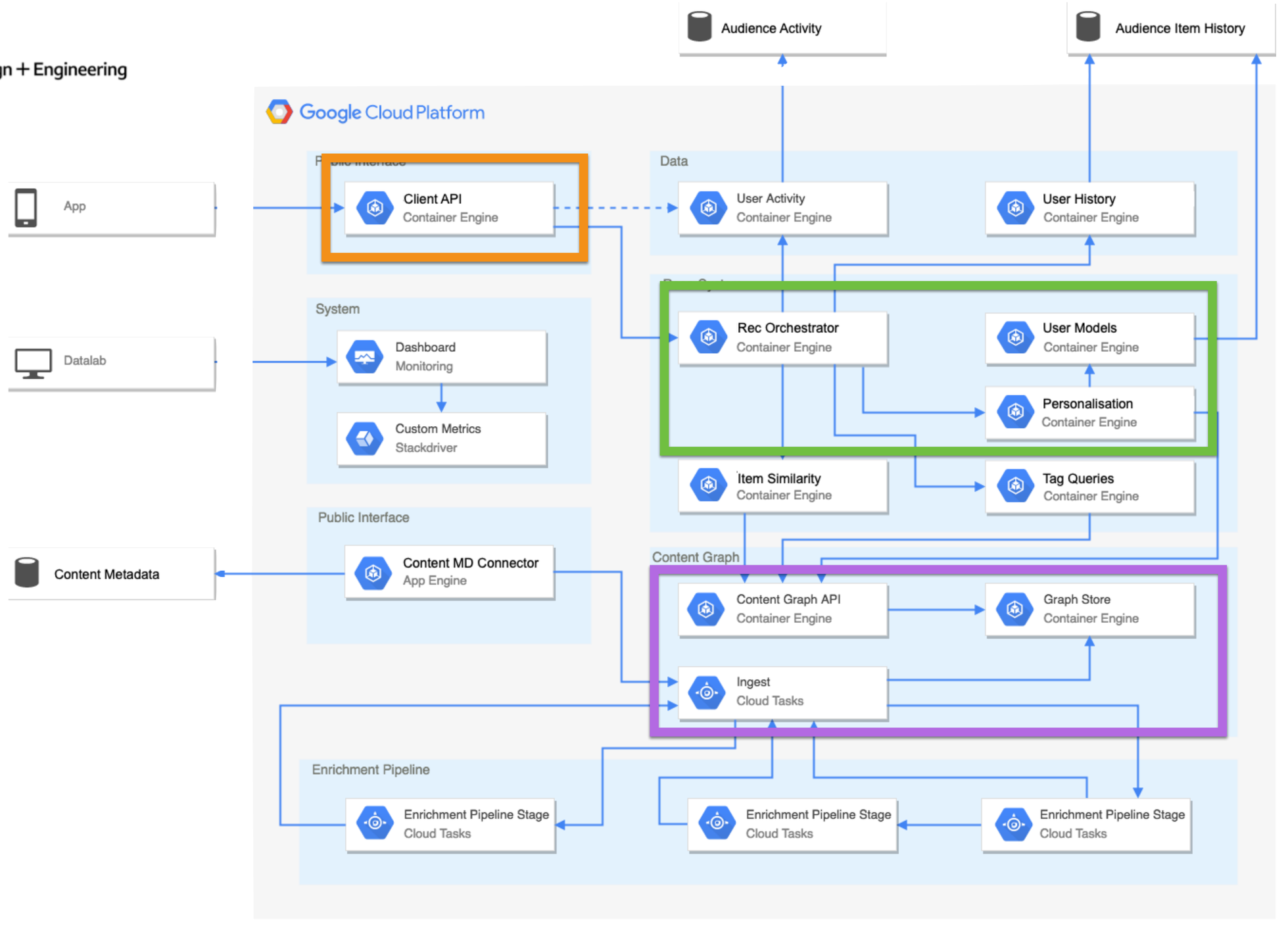
"We don't really know what the requirements are yet, so let's build something really flexible"

BBC+

OUR RESPONSE



OUR TECHNOLOGY CHOICES


Audience Activity
Audience Item History
Google Cloud Platform
App
Datalab
Content Metadata
Client API
Container Engine
Data
User Activity
Container Engine
User History
Container Engine
System
Dashboard
Monitoring
Custom Metrics
Stackdriver
Rec Orchestrator
Container Engine
User Models
Container Engine
Personalisation
Container Engine
Item Similarity
Container Engine
Tag Queries
Container Engine
Public Interface
Content MD Connector
App Engine
Content Graph
Content Graph API
Container Engine
Graph Store
Container Engine
Ingest
Cloud Tasks
Enrichment Pipeline
Enrichment Pipeline Stage
Cloud Tasks
Enrichment Pipeline Stage
Cloud Tasks
Enrichment Pipeline Stage
Cloud Tasks

Search or jump to... /

Pull requests Issues Marketplace Explore

BBC / TS&A / Datalab

Discussions Members 24 Teams 3 Repositories 80 Projects 0 Settings

Find a repository...

Add repository

Select all

bbc/connected-data Private updated 3 days ago Admin

bbc/connected-data-acceptance-tests Private updated on Apr 25 Admin

bbc/connected-data-amanita Private updated on May 28 Admin

bbc/connected-data-bramble Private updated on May 28 Admin

bbc/connected-data-bristlecone Private updated on May 28 Admin

bbc/connected-data-catopsis Private updated on May Admin

bramble_pb2_grpc.py

```python
# Generated by the gRPC Python protocol compiler plugin. DO NOT EDIT!
import grpc

import app.bramble_pb2 as bramble__pb2


class BrambleServiceStub(object):
  pass

  def __init__(self, channel):
    """Constructor.

    Args:
      channel: A grpc.Channel.
    """
    self.ListRecommendations = channel.unary_unary(
        '/bramble.BrambleService/ListRecommendations',
        request_serializer=bramble__pb2.ListRecommendationsRequest.SerializeToString,
        response_deserializer=bramble__pb2.Recommendations.FromString,
        )
    self.HealthCheck = channel.unary_unary(
        '/bramble.BrambleService/HealthCheck',
        request_serializer=bramble__pb2.Empty.SerializeToString,
        response_deserializer=bramble__pb2.Empty.FromString,
        )
    self.ListUserHistory = channel.unary_unary(
        '/bramble.BrambleService/ListUserHistory',
```

```python
clip_recommendations = bramble.ListRecommendations(user_id, media_type, time)
```

TECHNOLOGY CHOICES

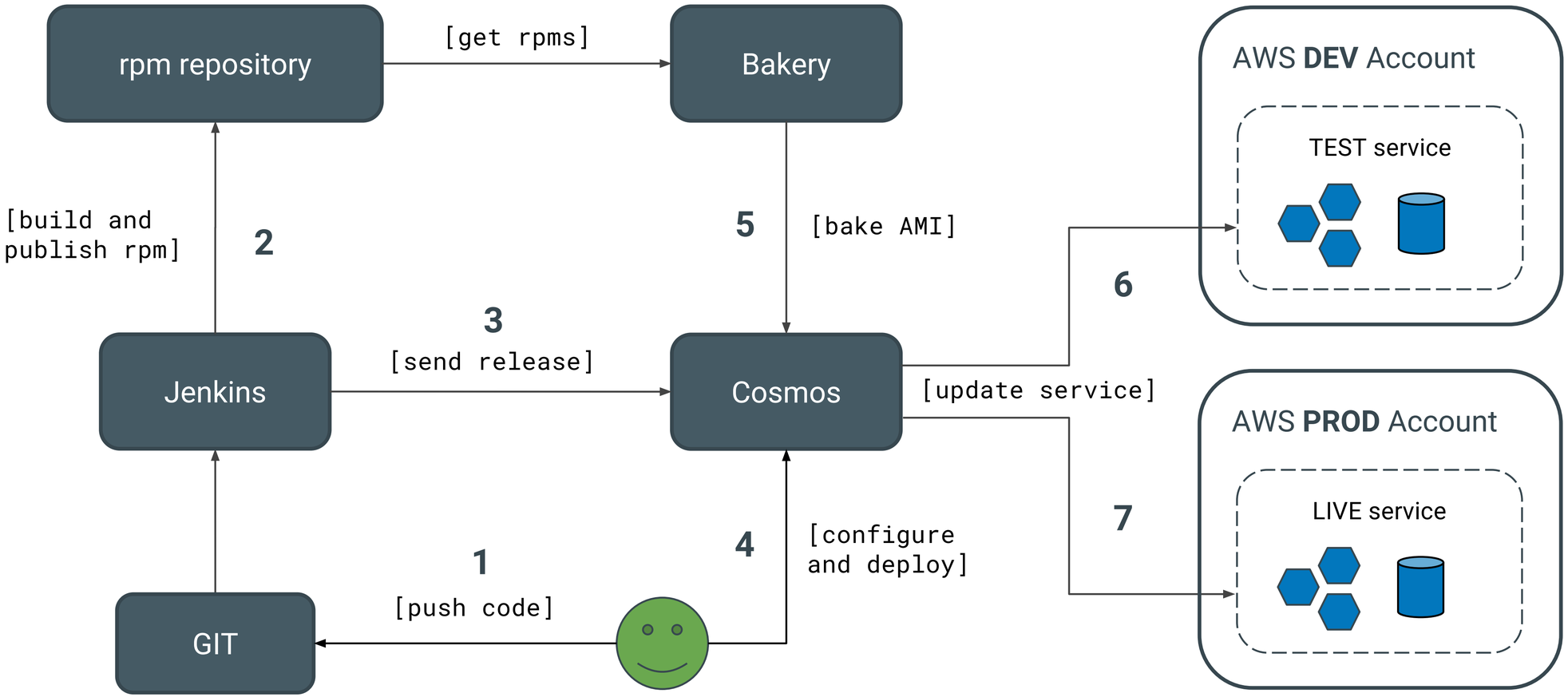
GRPC

HTTP (JSON)	gRPC (protobuf)
Every developer knows how to use (ish)	Developers need training
Tooling is everywhere	Tooling is difficult
Loads of python libraries!	gRPC library (un-googleable)
Slow?	~~Fast~~ Slow (in python)*
API changes are tricky	API changes backward-compatible

*https://performance-dot-grpc-testing.appspot.com/

TECHNOLOGY CHOICES

BBC "TRADITIONAL" SOFTWARE DEPLOYMENT



GOOGLE CLOUD PLATFORM

AI and machine learning
Text-to-Speech · Vision · Translation · More

API management
Apigee API Platform · Cloud Endpoints · More

Compute
Compute Engine · App Engine · More

Data analytics
BigQuery · Cloud Datalab · More

Databases
Cloud SQL · Cloud Datastore · More

Developer Tools
Container Registry · Cloud SDK · More

Hybrid and multi-cloud
Anthos · GKE On-Prem · Istio on GKE · More

Internet of Things
Cloud IoT Core · Edge TPU

Migration
Data Transfer · Transfer Appliance · More

Networking
DNS · CDN · Virtual Private Cloud · More

Security
Security Key Enforcement · Cloud IAM · More

Storage
Cloud Storage · Persistent Disk · More

MORE CLOUD PRODUCTS

G Suite
Gmail, Docs, Drive, Hangouts, and more

Google Maps Platform
Build with real-time, comprehensive data

Cloud Identity
Easily manage user identities

Chrome Enterprise
Get Chrome OS devices and browser

Android Enterprise
Intelligent devices, OS, and business apps

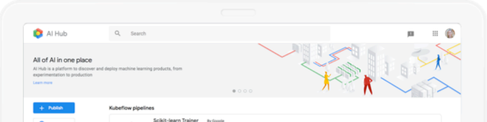
Hire by Google
Identify, evaluate, and hire better

See all products (100+)

AI Hub

Hosted repository of plug-and-play AI components

Google Cloud's AI Hub provides enterprise-grade sharing capabilities, including end-to-end AI pipelines and out-of-the-box algorithms, that let your organization privately

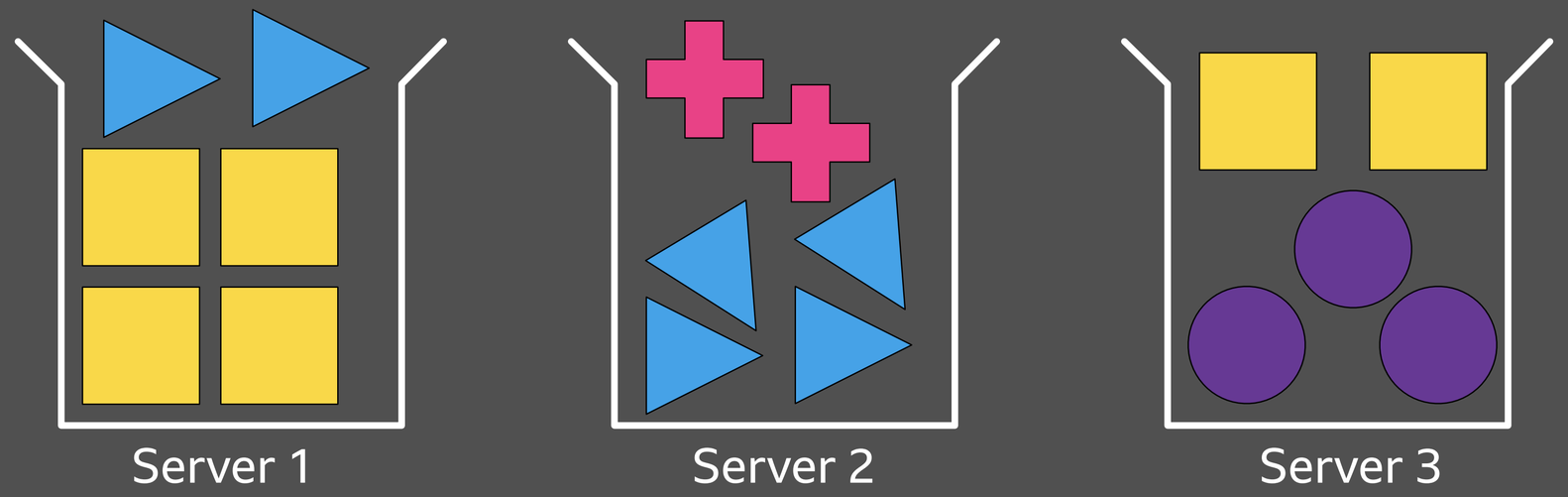


TECHNOLOGY CHOICES

DOCKER & KUBERNETES

```
1  FROM microservice-base:latest
2  MAINTAINER BBC Datalab <datalab@bbc.co.uk>
3
4  RUN pip3 install --upgrade pipenv && pipenv install --deploy --system --verbose
5
6  CMD /usr/local/bin/gunicorn --bind 0.0.0.0:5000 --workers 2 --access-logfile - app:server
```

```
some_service  master ● ? ↑1  kubectl apply -f service.yml
```



SPINNAKER Search Projects Applications Search What's New

datalabservice PIPELINES CLUSTERS TASKS LOAD BALANCERS SECURITY GROUPS CONFIG

Clusters

Edit multiple Show Instances with details Create Server Group

MY-K8S-ACCOUNT datalabservice-production 1 ▲ / 1 ▬ : 50%

DEFAULT

V006: cg-pov/datalab-service:1478ea85412204708419bb99f419e466584938c0 1 ▲ : 100%

V005: cg-pov/datalab-service:545246b8a203e6dbaa632033450923b84005dd91 1 ▬ : 0%

MY-K8S-ACCOUNT datalabservice-stage 1 ▲ : 100%

DEFAULT

V008: cg-pov/datalab-service:1478ea85412204708419bb99f419e466584938c0 1 ▲ : 100%

datalabservice-live

Load Balancer Actions

LOAD BALANCER DETAILS

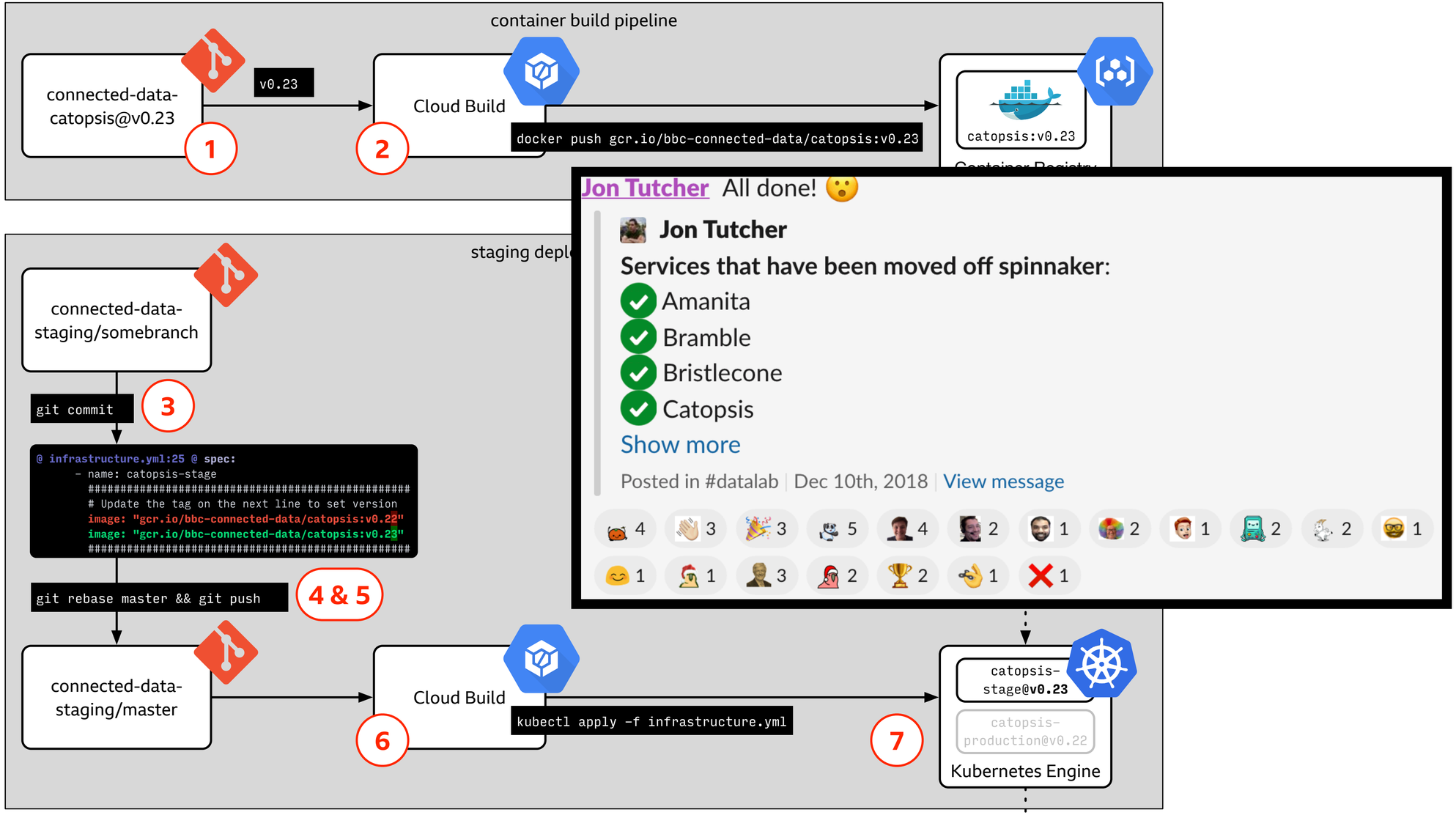
Created 2017-12-15 07:41:42 PST
In MY-K8S-ACCOUNT
Namespace default
Kind Service
YAML Show YAML
Kube UI datalabservice-live

Server Groups datalabservice-production-v006
datalabservice-production-v005

Service Type
LoadBalancer

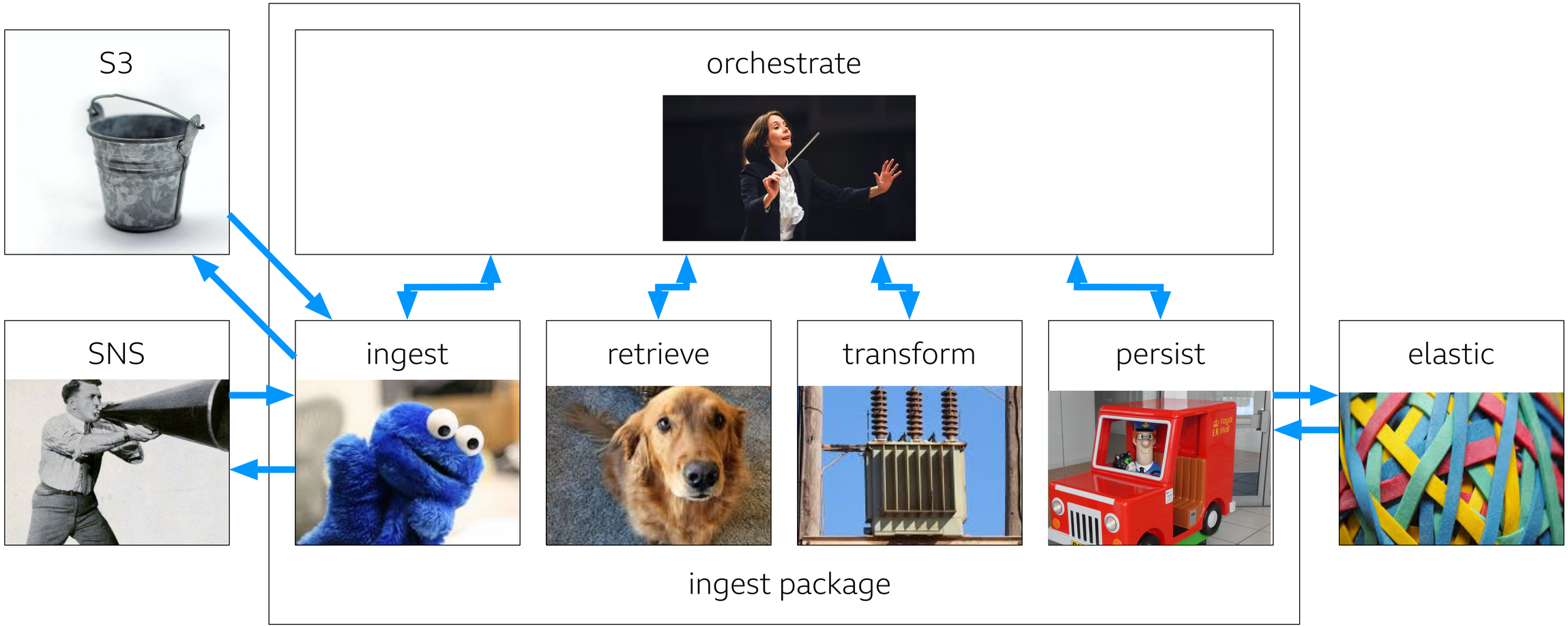
Session Affinity
None

Internal DNS Name
datalabservice-live.default.svc.cluster.local


container build pipeline
connected-data-catopsis@v0.23
v0.23
1
Cloud Build
2
docker push gcr.io/bbc-connected-data/catopsis:v0.23
catopsis:v0.23
Jon Tutcher All done! 😮
Jon Tutcher
Services that have been moved off spinnaker:
Amanita
Bramble
Bristlecone
Catopsis
Show more
Posted in #datalab | Dec 10th, 2018 | View message
staging depl
connected-data-staging/somebranch
git commit
3
@ infrastructure.yml:25 @ spec:
- name: catopsis-stage
Update the tag on the next line to set version
image: "gcr.io/bbc-connected-data/catopsis:v0.22"
image: "gcr.io/bbc-connected-data/catopsis:v0.23"
git rebase master && git push
4 & 5
connected-data-staging/master
Cloud Build
6
kubectl apply -f infrastructure.yml
7
catopsis-stage@v0.23
catopsis-production@v0.22
Kubernetes Engine

OVERALL TEAM EFFECTS

- Pace dropped
- Low confidence in our code
- Data science dev slowed
- Bugs compounded
- Team morale dropped (until we started fixing!)


S3
orchestrate
SNS
ingest
retrieve
transform
persist
elastic
ingest package

File Edit View Run Kernel Tabs Settings Help

SEARCH

CONSOLE

Change Kernel...
Clear Console Cells
Close and Shutdown...
Insert Line Break
Interrupt Kernel
New Console
Restart Kernel...
Run Cell (forced)
Run Cell (unforced)
Show All Kernel Activity

FILE OPERATIONS

✓ Autosave Documents
Close All
Close Other Tabs
Close Python File ^ Q
Close Tabs to Right
New View for Python File
Open From Path...
Reload Python File from Disk
Revert Python File to Checkpoint
Save Python File ⌘ S
Save Python File As... ⇧ ⌘ S

HELP

JupyterLab Reference
Launch Classic Notebook
Markdown Reference

Launcher | locustfile.py

```python
import json

import numpy as np

import gevent.pool
from locust import TaskSet, Locust, task, HttpLocust

class HitApiEndpointTasks(TaskSet):

    """
    Define a set of tasks to run *per user/locust*.

    Tasks
        genres: a locust task that concurrently requests a number of videos, `LIMIT`,
            for a number of genres, `N_GENRE_REQUESTS`. Genres are randomly sampled.
    """

    @task
    def predict_flask(self):
        LENGTH_OF_VECTOR=50
        NUMBER_OF_VECTORS=2

        vectors = []
        for  vector in range(NUMBER_OF_VECTORS):
            topic_percentages = np.random.random(LENGTH_OF_VECTOR)
            #the numbers in the vector should sum to 1
            topic_percentages /= topic_percentages.sum()
            vectors.append(list(topic_percentages))


        locust_response = self.client.post('/predict', verify=False, data=str(vectors))
        print("Response status code:", locust_response.status_code)
        print("Response content:", locust_response.text)


class HitTfServeEndpointTasks(TaskSet):
    @task
    def predict_tfserving(self):
        LENGTH_OF_VECTOR=75
```

TECHNOLOGY CHOICES

OTHER GOOD IDEAS

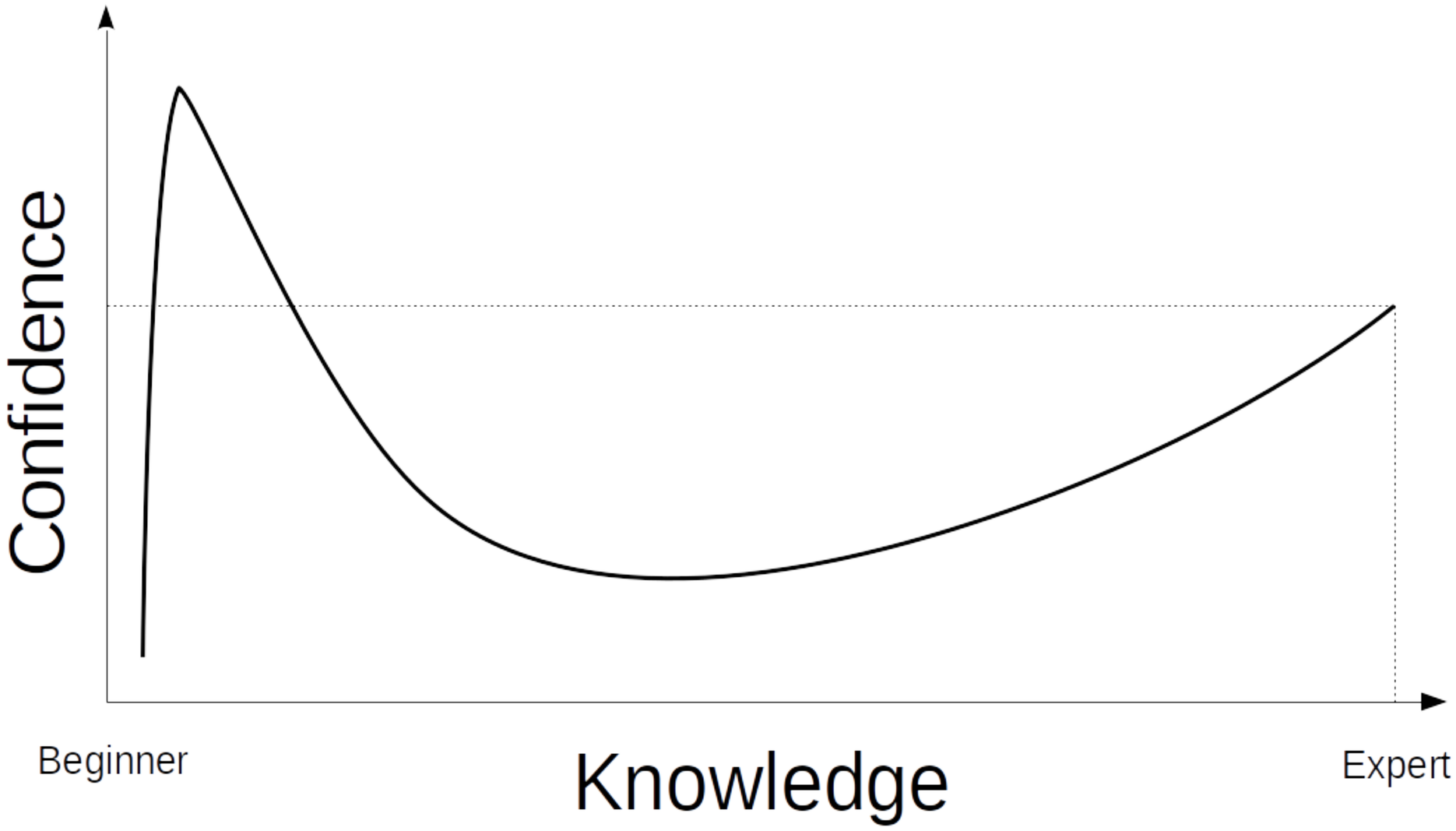
- Elasticsearch
- Managed Logging (mostly)
- Managed Training (Google ML Engine) (mostly)
- Committing to Tensorflow (for now)

TECHNOLOGY CHOICES

LESSONS LEARNT

- Decision making in new teams
- Over-engineering is easier than doing research
- Selection bias in press / meetups

- Python = no hassle
- Kubernetes = a keeper (for larger projects)



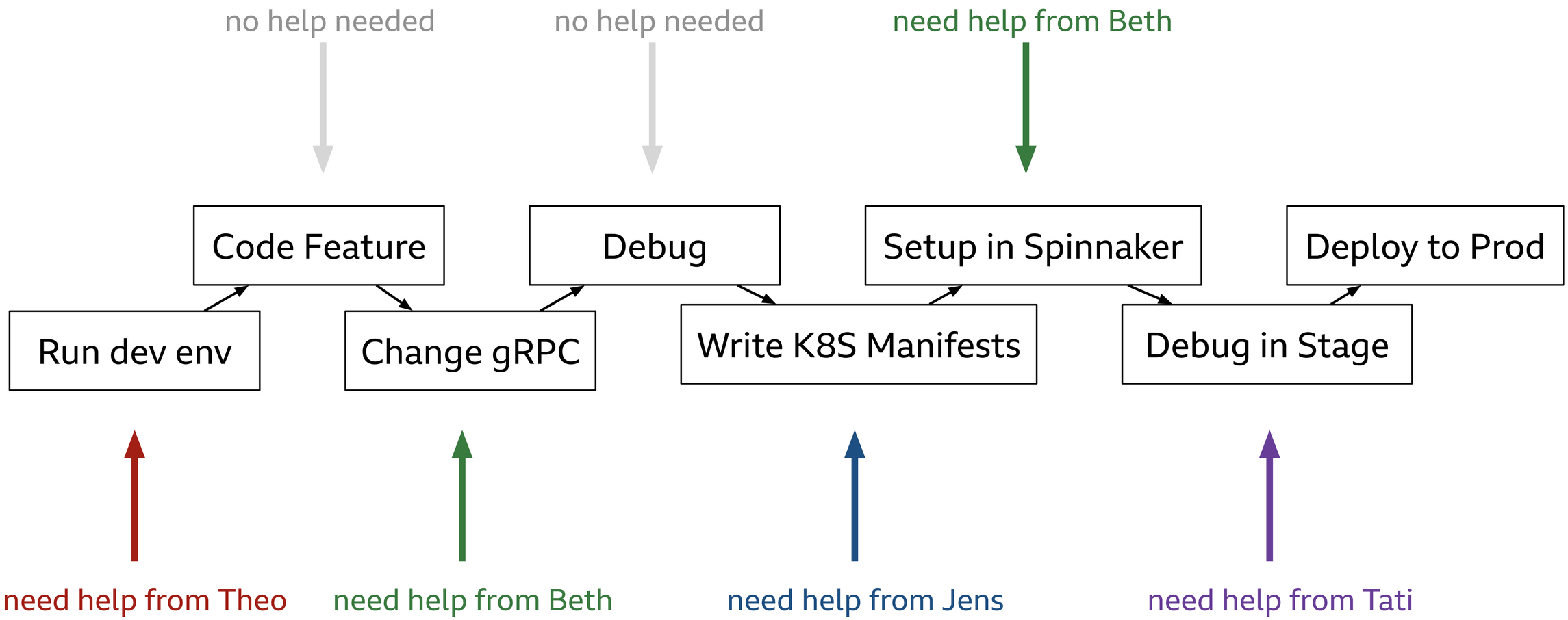
https://commons.wikimedia.org/wiki/File:Dunning-kruger.png

"USE BORING TECHNOLOGY"

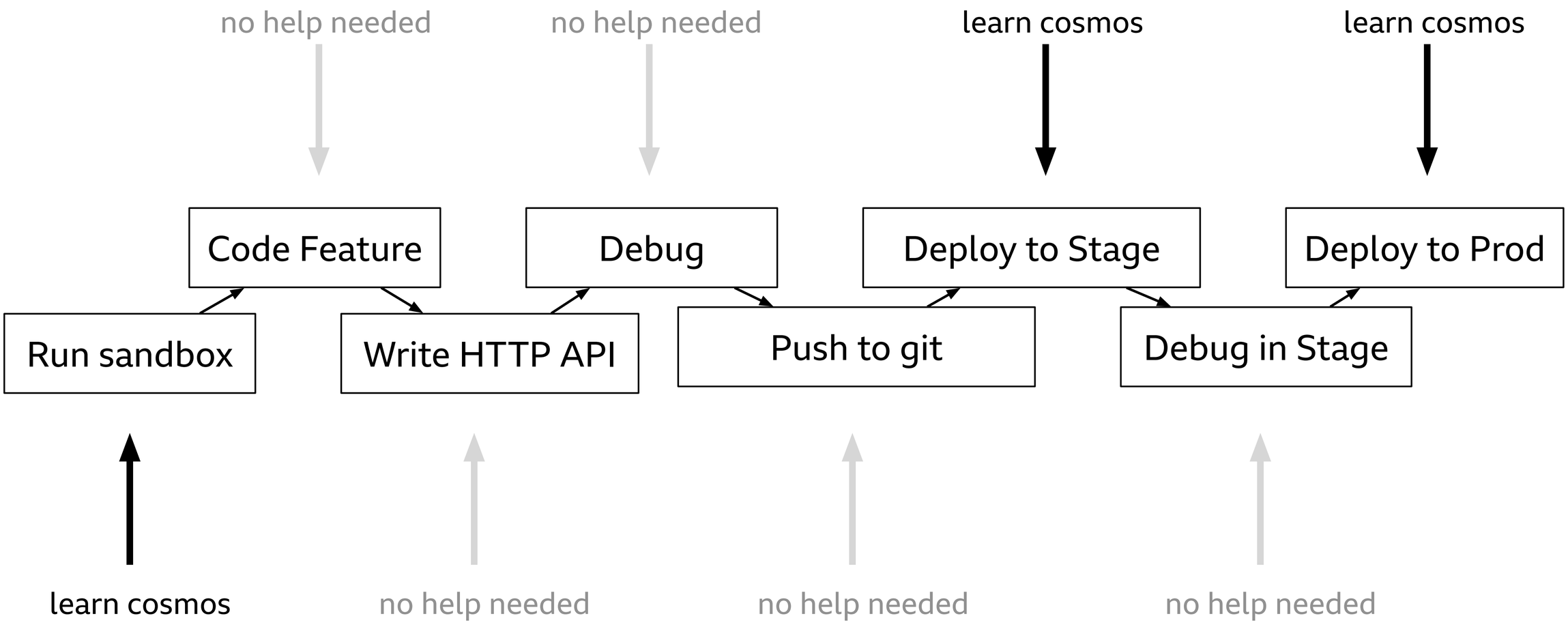
"The grim paradox of this law of software is that you should probably be using the tool that you hate the most. You hate it because you know the most about it."

- Dan McKinley, http://boringtechnology.club/

ANATOMY OF TASK BLOCKING



ANATOMY OF TASK BLOCKING



*cosmos = BBC's cloud deployment platform

```python
class Technology:
    def __init__(self, name, maturity, familiarity, support, maintenance_cost, benefit):
        self.name = name
        self.maturity = maturity
        self.support = support
        self.familiarity = familiarity
        self.maintenance_cost = maintenance_cost
        self.benefit = benefit

    @property
    def pace_cost(self):
        risk = (1 - self.maturity) + (1 - self.support)
        return risk * (1 - self.familiarity)

    @property
    def total_cost(self):
        benefits = self.benefit
        risks = self.pace_cost + self.maintenance_cost
        return max(risks - benefits, 0)

technologies = [
    Technology("Spinnaker", maturity=0.1, familiarity=0.2, support=0.4, maintenance_cost=0.7, benefit=0.5),
    Technology("Postgres", maturity=1.0, familiarity=0.8, support=1.0, maintenance_cost=0.5, benefit=0.7),
    Technology("Hosted SQL", maturity=0.7, familiarity=0.5, support=0.7, maintenance_cost=0.2, benefit=0.7),
    Technology("Airflow", maturity=0.2, familiarity=0.5, support=0.4, maintenance_cost=0.4, benefit=0.6)
]

# Model cost of adoption
for tech in technologies:
    print(f"{tech.name}: pace cost: {tech.pace_cost:.2f}, total cost: {tech.total_cost:.2f}")
```

```
Spinnaker: pace cost: 1.20, total cost: 1.40
Postgres: pace cost: 0.00, total cost: 0.00
Hosted SQL: pace cost: 0.30, total cost: 0.00
Airflow: pace cost: 0.70, total cost: 0.50
```

BBC+

THE NEXT CHALLENGE

"How can we make model exploration and creation as automated as possible, whilst tracking provenance of data and code"?

WHAT NEXT?

THE KNEEJERK REACTION

mlflow™

APACHE kafka®

DASK

TensorFlow Extended

SELDON

Kubeflow

papermill

ML WORKFLOW TOOLS

WHAT'S NEXT?

```
new_techs = [Technology("Luigi", 0.2, 0.1, 0.3, 0.7, 0.7),
             Technology("MLFlow", 0.3, 0.2, 0.3, 0.7, 0.7),
             Technology("Dask", 0.4, 0.3, 0.5, 0.6, 0.7),
             Technology("Kafka", 0.6, 0.7, 0.7, 0.7, 0.7),
             Technology("Beam", 0.4, 0.5, 0.4, 0.7, 0.8),
             Technology("Jenkins", 0.9, 0.7, 0.9, 0.2, 0.6)]
```

```
Luigi: pace cost: 1.35, total cost: 1.35
MLFlow: pace cost: 1.12, total cost: 1.12
Dask: pace cost: 0.77, total cost: 0.67
Kafka: pace cost: 0.21, total cost: 0.21
Beam: pace cost: 0.60, total cost: 0.50
Jenkins: pace cost: 0.06, total cost: 0.00
```

FINAL THOUGHTS

- Fit your problem to existing tech (if poss)
- Avoid sunk cost fallacy
- Experiment, but one-at-a-time
- What's right for Google isn't right for you

THANKS!

@jontutcher

Come and work with us!

https://findouthow.datalab.rocks/