BUILDING DATA WORKFLOWS

breuninger

with

LUIGI & KUBERNETES

EuroPython 2019, Basel

INTRO @CHHANTYAL

breuninger

Nar Kumar Chhantyal

- Data Lake @ Breuninger.com
- Python/Luigi with Kubernetes on Google Cloud
- Web Dev in past life (Flask/Django/NodeJS)
- Twitter/Github: @chhantyal
- Web: http://chhantyal.net

WHAT IS LUIGI?

breuninger

- Workflow/pipeline tool for batch jobs
- Open sourced by Spotify Engineering
- Written entirely in Python. Jobs are just normal Python code
- Lightweight, comes with Web UI
- Has tons of contrib packages eg. Hadoop, BigQuery, AWS
- Has no built in scheduler, usually crontab is used

EXAMPLE USECASE

breuninger

Daily Sales Report

Create a daily revenue report from sales transactions.

We need do few things first to build final report:

- Dump sales data from prod database
- Ingest into analytics database
- Run aggregation & update dashboard

EXAMPLE USECASE

breuninger

Daily Sales Report

I will just write modular Python script, what could possibly go wrong?

1. 0 10 * * * dump_sales_data.py
2. 0 11 * * * ingest_to_analyticsdb.py
3. 0 12 * * * aggregate_data.py
4. Profit? 😎

EXAMPLE USECASE

breuninger

Daily Sales Report

Few issues:

1. What happens when first one fails?
2. What if first one takes longer than one hour?
3. What if you have to do same thing for last five days?
4. How do I see if these jobs ran successfully or not?
5. What happens if job somehow runs twice? Duplicate data?

EXAMPLE USECASE

breuninger

Daily Sales Report

- Luigi implimentation
- Source code: https://github.com/chhantyal/luigi-kubernetes
- Run from CLI: luigi --module example SalesReport --date=2019-07-11

RUNING LUIGI WITH CRONTAB

Luigi has no built-in scheduler. Usually, crontab is used:

- 0 08 * * * luigi --module example SalesReport --date=2019-07-11

+

CRONTAB

LUIGI ON KUBERNETES

breuninger

Luigi having no built-in scheduler is blessing in disguise.

+

Kubernetes Cronjob

KUBERNETES (CRON)JOB

breuninger

A Job creates one or more Pods to do specific task. It ensures the pods' successful completion and reschedules them in case of failure (aka. run to complation).

A *Cron Job* creates Jobs on a time-based schedule.

RUNNING ON KUBERNETES

breuninger

Daily Sales Report

- Run on Kubernetes (Minikube)
  - Deploy Luigid
  - Build Docker images & upload to registry
  - Deploy pipeline on K8S
- Cronjob → Job → Pod
- Source code: https://github.com/chhantyal/luigi-kubernetes
- Docker images: https://hub.docker.com/u/chhantyal

SUMMARY

breuninger

Luigi being lightweight, it makes great tool to containerize and run on Kubernates cluster.

As a result, you can manage complex batch processes and scale them seamlessly on demand.

Luigi

- Workflow managment
- Dependency resolution
- Easy testing & containerization

Kubernetes

- Horizontal scaling
- Flexible deployment
- Continuous integration & delivery

DATA TEAM @ BREUNINGER IS HIRING!

breuninger

Contact: kumar.chhantyal@breuninger.de | twitter.com/chhantyal

- Data (big & small)
- Python 💙
- Docker/Kubernetes
- Google Cloud
- Table tennis 🏓 / running 🏃 / biking 🚴 / cakes ✨🍰✨
- Cool team 😎
- Stuttgart, Germany (ca. 2h train ride from Basel)

breuninger

THANKS FOR JOINING ME!

QUESTIONS?

Do you use Python for Data Engineering? Happy to chat about it ☺

Docker images: https://hub.docker.com/u/chhantyal
Source code: https://github.com/chhantyal/luigi-kubernetes