DATA COUNCIL SINGAPORE 2019

ARGO: KUBERNETES NATIVE WORKFLOWS & PIPELINES

GREG ROODT - DATA ENGINEERING LEAD

AGENDA

INTRO TO CANVA

WHAT IS A WORKFLOW?

ARGO VS AIRFLOW

ARGO DEMO

Q & A

I.

INTRO TO CANVA

”

STARTUP FROM SYDNEY

“

EMPOWERING THE WORLD TO DESIGN

II.

WHAT IS A WORKFLOW?

OXFORD ENGLISH DICTIONARY

NOUN

The sequence of industrial, administrative, or other processes through which a piece of work passes from initiation to completion.

https://www.lexico.com/en/definition/workflow

WIKI
PEDIA

A workflow consists of an orchestrated and repeatable pattern of activity...

https://en.wikipedia.org/wiki/Workflow

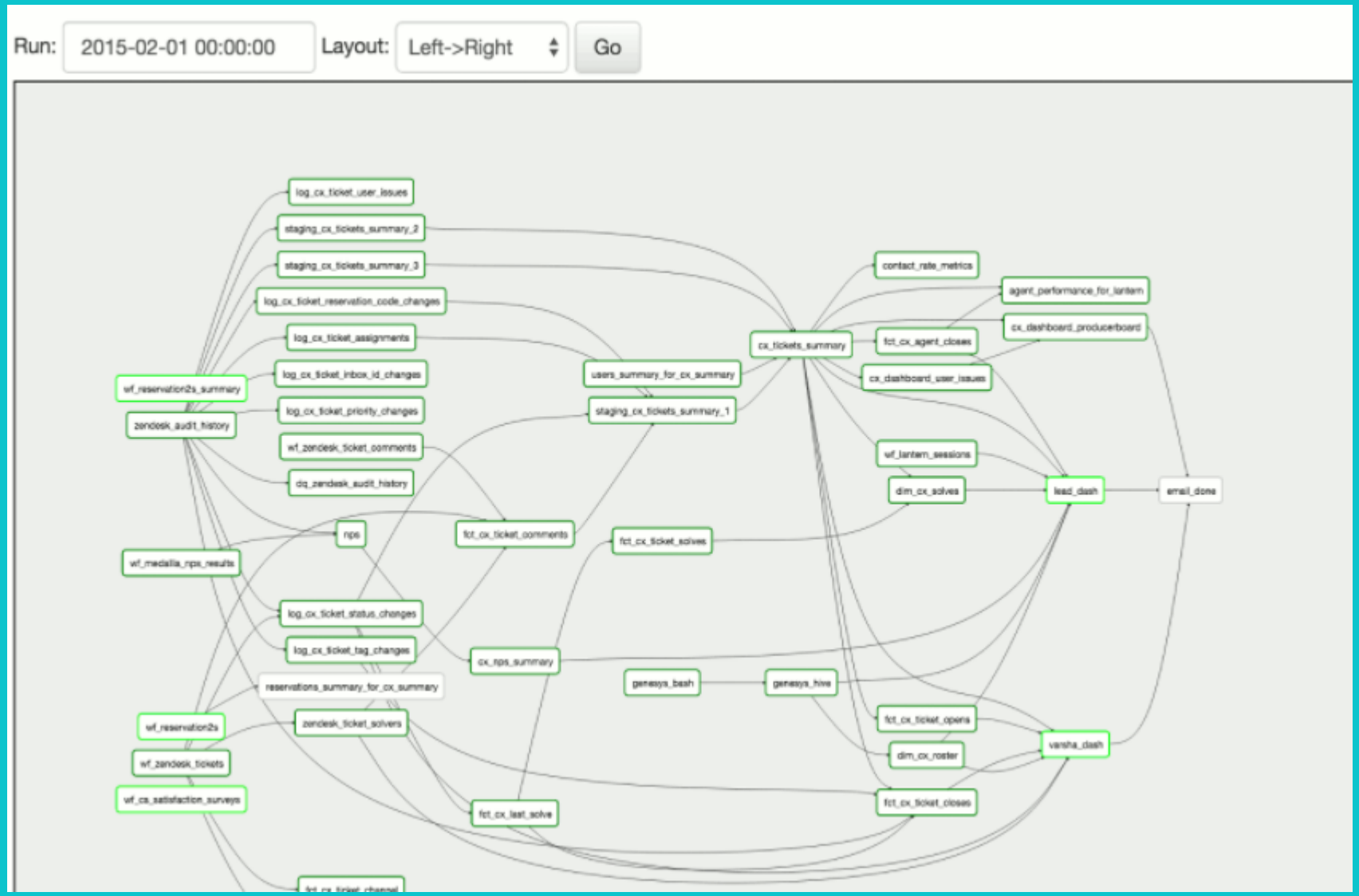
DIRECTED ACYCLIC GRAPH (DAG)

- We typically mean a DAG
- Steps in a DAG are "tasks"
- Each "task" is some discrete amount of work
- All the "tasks" together make up some larger piece of work
- Examples: ETL jobs, Spark jobs, Machine Learning jobs, CI jobs, deployment jobs etc.

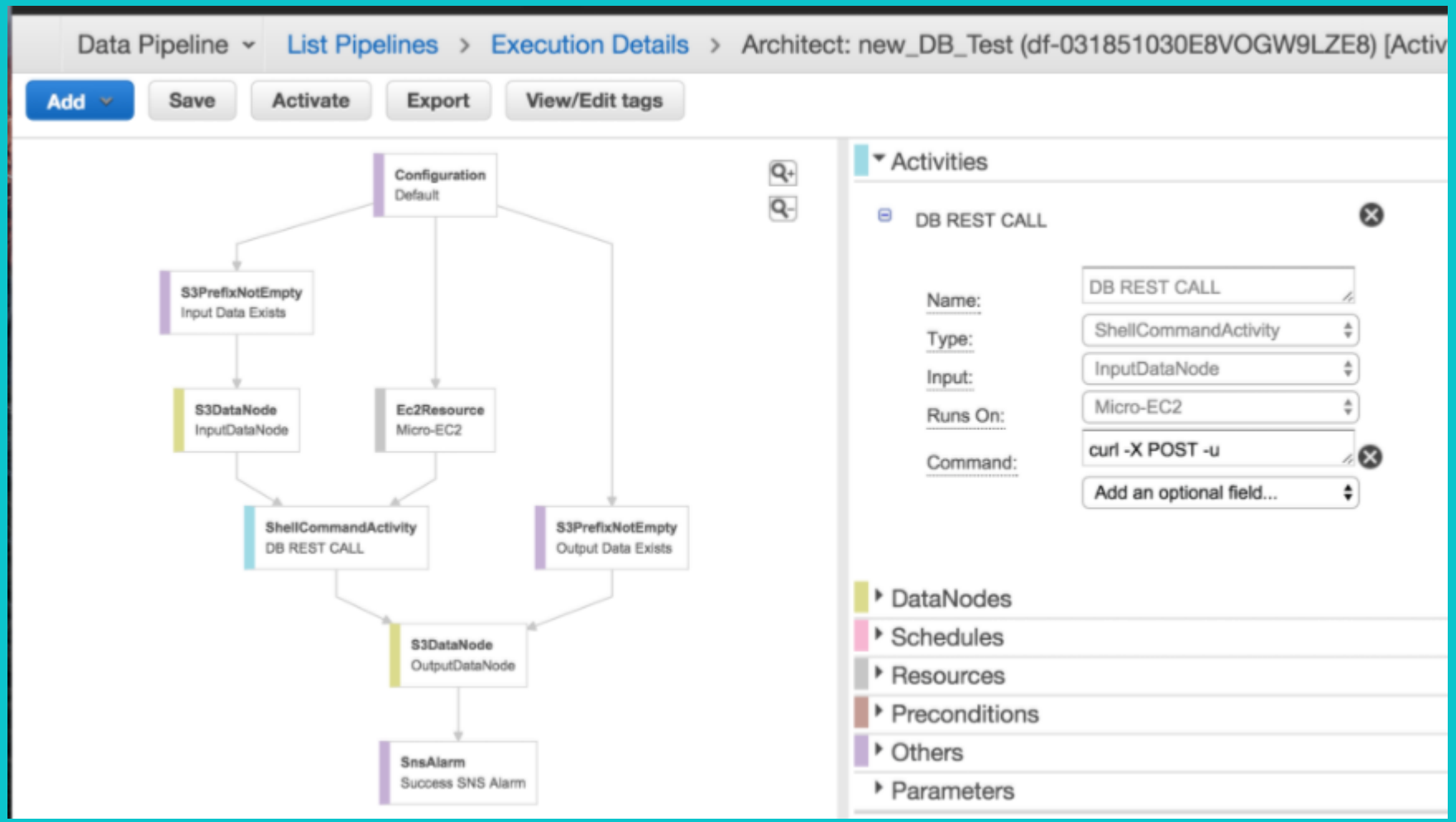
WHERE TO START?

- So many different tools out there
- Oozie, Luigi, NiFi, Azkaban...
- The most popular one at the moment seems to be Apache Airflow

AIRFLOW DAG


Run:
2015-02-01 00:00:00
Layout:
Left->Right
Go
log_cx_ticket_user_issues
staging_cx_tickets_summary_2
staging_cx_tickets_summary_3
log_cx_ticket_reservation_code_changes
log_cx_ticket_assignments
log_cx_ticket_inbox_id_changes
log_cx_ticket_priority_changes
wf_zendesk_ticket_comments
dq_zendesk_audit_history
wf_reservation2s_summary
zendesk_audit_history
nps
wf_medallia_nps_results
log_cx_ticket_status_changes
log_cx_ticket_tag_changes
reservations_summary_for_cx_summary
wf_reservation2s
zendesk_ticket_solvers
wf_zendesk_tickets
wf_cs_satisfaction_surveys
fct_cx_ticket_comments
users_summary_for_cx_summary
staging_cx_tickets_summary_1
fct_cx_ticket_solves
cx_nps_summary
fct_cx_last_solve
genesys_bash
genesys_hive
cx_tickets_summary
contact_rate_metrics
fct_cx_agent_closes
cx_dashboard_user_issues
agent_performance_for_lantern
cx_dashboard_producerboard
wf_lantern_sessions
dim_cx_solves
lead_dash
email_done
fct_cx_ticket_opens
dim_cx_roster
vansha_dash
fct_cx_ticket_closes

AWS DATAPIPELINE



ARGO DAG


kubeflow-presu...
checkout
setup
deploy-tf-serving
deploy-tf-servin...
create-pr-symlink
test-tf-serving
test-tf-serving-g...
kubeflow-presu...
teardown

III.

ARGO VS AIRFLOW

FIRST A STORY

- We were a very small team at this point...
- First there was AWS Datapipeline
- And it was good
- Complicated tool, but it's a complicated problem and it was reliable

TIME PASSES

- Time went on. Team grew in ambition
- Started running into limitations with Datapipeline
- Frustrations grew

SOME ISSUES

- Limited EC2 instance types (less true now)
- Custom software is awkward
- Scheduling and retries is awkward, we built our own workarounds
- Not very well known tool (no community)

ARGO AIRFLOW

- We decided to compare alternatives
- Airflow was the obvious choice
- Argo was a new project that ticked a lot of boxes on paper
- Did a comprehensive proof-of-concept to compare

AIRFLOW

- Originated at Airbnb
- Python based operators & DAG
- Huge community
- Open source (Host yourself)
- Commercial SaaS offerings available too

https://airflow.apache.org/

ARGO

- Originated at Intuit (Applatix)
- K8s Native (CRD+Controller)
- Written in Golang
- DAG of Containers (Any language)
- Growing community
- Host your own on K8s

https://argoproj.github.io/

PROOF OF CONCEPT

- Decided on simple, realistic example
- Setup both on AWS
- Came up with a few evaluation criteria
- Got to work!

EVALUATION CRITERIA

- Suitability for Canva (mostly Spark) workloads
- Ease of setup and operation
- Good workflow deployment
- Logging
- Live UI updates
- Timeouts & retries
- Reproducibility

THINGS WE WERE NOT LOOKING FOR

- Visual UI for editing DAGs
- Extremely Dynamic DAGs
- Python-only
- Updating a DAG after workflow submission

GOOD PARTS?

- Both tools are very capable
- Both tools could support Canva workloads
- Both tools offer live UI & logging updates
- Both tools offer timeouts and retries

ISSUES WITH AIRFLOW

- Deployment of DAGs is clumsy (poll git, rsync?)
- Mutable DAGs
- DAG deployed is not necessarily the DAG that runs
- Installation of components fairly complex
- A lot of excitement around k8s executor, so why not use k8s directly?

ISSUES WITH ARGO

- Lots of YAML
- Requires k8s
- Simpler, less powerful UI
- Early project

THE DECISION?

- Both projects very capable
- Airflow lacks solid deployment story
- Chose Argo

IV.

ARGO DEMO

THANK YOU

WE'RE HIRING!
CANVA.COM/CAREERS