Airflow

Arun

May 27, 2019

Airflow

What is Airflow?

What is Airflow?

- It's a workflow manager

What is Airflow?

- It's a workflow manager
- We can also call it a fancy Scheduler

What is Airflow?

- It’s a workflow manager
- We can also call it a fancy Scheduler
- It can be a easy backend job Scheduler in a distributed system
- Ref, this blog

Other tools

Other tools

- Oozie - XML code

Other tools

- Oozie - XML code
- Pinball (pinterest) - no community interest

Other tools

- Oozie - XML code
- Pinball (pinterest) - no community interest
- Luigi (Spotify) - Uses HDFS, no alerts or monitoring

Other tools

- Oozie - XML code
- Pinball (pinterest) - no community interest
- Luigi (Spotify) - Uses HDFS, no alerts or monitoring
- Azkaban (LinkedIn) - Uses custom config file for workflow setups

Airflow

Airflow

- Python code base + for workflow definitions

Airflow

- Python code base + for workflow definitions
- Trigger rules

Airflow

- Python code base + for workflow definitions
- Trigger rules
- Xcoms

Airflow

- Python code base + for workflow definitions
- Trigger rules
- Xcoms
- Cool UI & CLI

Airflow

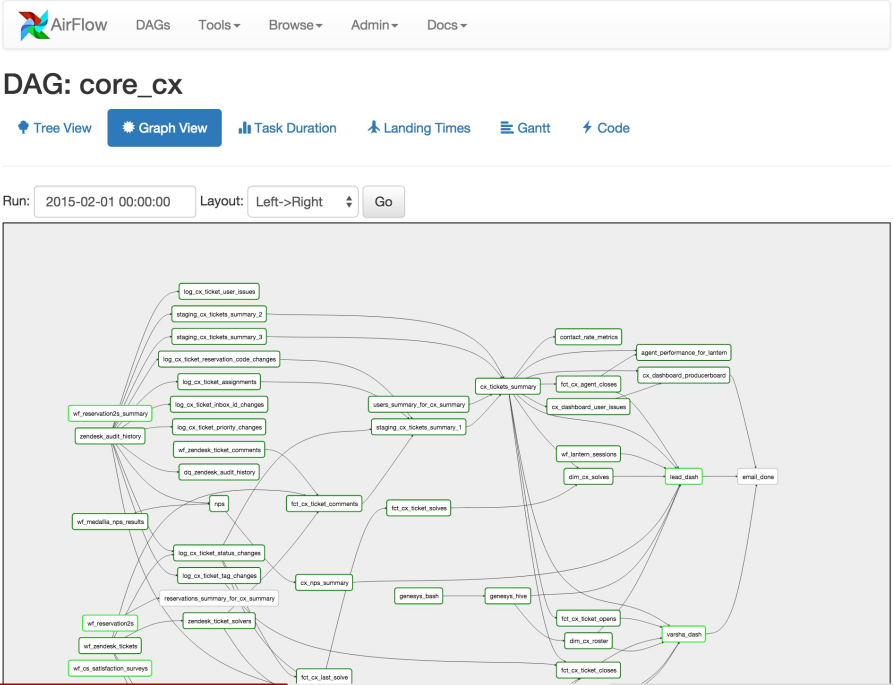
- Python code base + for workflow definitions
- Trigger rules
- Xcoms
- Cool UI & CLI
- Queues & Pools

Airflow

- Python code base + for workflow definitions
- Trigger rules
- Xcoms
- Cool UI & CLI
- Queues & Pools
- Zombie cleanup

Airflow

- Python code base + for workflow definitions
- Trigger rules
- Xcoms
- Cool UI & CLI
- Queues & Pools
- Zombie cleanup
- Large community



DAG: core_cx

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
wf_reservation2s_summary
zendesk_audit_history
log_cx_ticket_priority_changes
wf_zendesk_ticket_comments
dq_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
users_summary_for_cx_summary
staging_cx_tickets_summary_1
fct_cx_ticket_comments
fct_cx_ticket_solves
cx_nps_summary
genesys_bash
genesys_hive
fct_cx_last_solve
cx_tickets_summary
contact_rate_metrics
agent_performance_for_lantern
cx_dashboard_producerboard
fct_cx_agent_closes
cx_dashboard_user_issues
wf_lantern_sessions
dim_cx_solves
lead_dash
email_done
fct_cx_ticket_opens
varsha_dash
dim_cx_roster
fct_cx_ticket_closes

Airflow - Executors

- Sequential
- Local (parallel)
- Celery
- Mesos
- Kubernetes (still in alpha)

Our cluster

- Single node (8 core, 20 GB node)
- Local Executor (with 24 parallel schedulable tasks)
- 3000+ tasks triggred a day
- Code auto deployed from master to prod

Tasks

- BQ job
- Run spark jobs daily for ML
- All batched jobs in Google Dataflow
- Scheduling backend jobs
- All errors to SLACK
- Critical DAG failures will invoke Victorops issue