Airflow at WePay

Chris Riccomini · June 14, 2016

Who?

Chris Riccomini
Engineer at WePay
Working on infrastructure (mostly)
Formerly LinkedIn, PayPal

Who?

Goal

- What do we use Airflow for?
- How do we operate Airflow?

Usage

- ETL
- Reporting
- Monitoring
- Machine learning
- CRON replacement

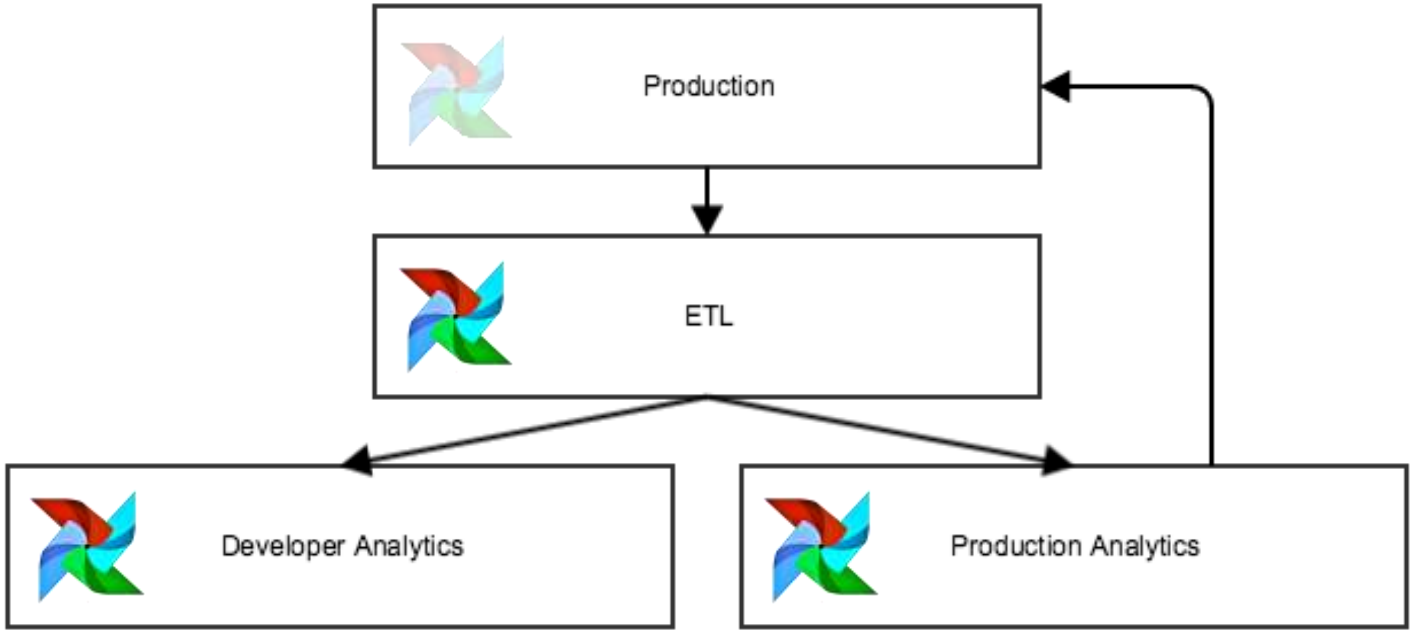
Usage

- ETL
- Reporting
- Monitoring
- Machine learning
- CRON replacement

Usage

- 350 DAGs
- 7,000 DAG runs per-day
- 20,000 task instances/day

Environments



Airflow deployment

- Google cloud platform
- One n1-highcpu-32 machine (32 cores, 28G mem)
- CloudSQL hosted MySQL 5.6 (250GB, 12GB used)
- Supervisord
- Icinga2 (investigating Sensu)

Airflow deployment

DataInfra / airflow

Watch 3 | Star 0 | Fork 2

Code | Issues 0 | Pull requests 1 | Pulse | Graphs

A clone of https://github.com/airbnb/airflow with custom patches applied. This allows us to use our patches before they're committed to AirBNB's Airflow repo.

3,059 commits | 5 branches | 58 releases | 4 contributors

Branch: master | New pull request | New file | Find file | HTTPS | https://github.devops.w | Download ZIP

Switch branches/tags

Filter branches/tags

Branches | Tags

- 1.6.2-wepay
- 1.7.0-wepay
- 1.7.1.2-wepay
- ✓ master
- wepay-master

132/master ··· Latest commit fcba070 on Apr 17

	ned PULL_REQUEST_TEMPLATE	2 months ago
	n the scheduler: num_runs used where runs intended	2 months ago
	hook and operator	11 months ago
	et/set variables in the CLI	2 months ago
	cessing support to the scheduler	2 months ago
	quest #1376 from bolkedebruin/multiprocessing_scheduler	2 months ago
	rom coverage report	6 months ago
.coveralls.yml	[hotfix] removing repo_token from .coveralls.yml	3 months ago
.gitignore	This patch adds license checking for Airflow. For now it will store a…	3 months ago
.landscape.yml	Linting	3 months ago
.rat-excludes	Add support for zipped dags	2 months ago

Airflow deployment

```
pip install git+https://git@our-wepay-repo.com/DataInfra/airflow.git@1.7.1.2-wepay#egg=airflow[gcp_api,mysql,crypto]==1.7.1.2+wepay4
```

Airflow scheduler

- Single scheduler on same machine as webserver

```
executor = LocalExecutor
parallelism = 64
dag_concurrency = 64
max_active_runs_per_dag = 16
```

Airflow logs

- /var/log/airflow
- Remote logger points to Google cloud storage
- Experimenting with ELK

Airflow connections

Connections

List (10) Create With selected

	Conn Id	Conn Type	Host	Port	Is Encrypted	Is Extra Encrypted	Extra
	airflow	mysql			✓		+
	db_monolith_fraud	mysql			✓		+
	db_monolith_log	mysql			✓		+
	db_monolith_products	mysql			✓		+
	db_monolith_wepay	mysql			✓		+
	gcp_api	google_cloud_platform			⊖	✓	+
	gcp_bi	google_cloud_platform			⊖	✓	+
	gcp_core	google_cloud_platform			⊖	✓	+
	gcp_di	google_cloud_platform			⊖	✓	+
	gcp_risk	google_cloud_platform			⊖	✓	+

Airflow security

- Active directory
- LDAP Airflow backend
- Disabled admin and data profiler tabs

DAG development

DataInfra / **airflow-dags**

Watch 2 | Star 0 | Fork 5

Code | Issues 0 | Pull requests 1 | Pulse | Graphs

Repository for Airflow DAGs.

508 commits | **9** branches | **0** releases | **4** contributors

Branch: **master** | New pull request | New file | Find file | HTTPS | https://github.devops.w | Download ZIP

ChrisRiccomini DI-319 Add wepay.reports to ETL pipeline (#234) — Latest commit 4495036 4 days ago

dags	DI-319 Add wepay.reports to ETL pipeline (#234)	4 days ago
dockerfiles	Update Dockerfile per David Clarke's recommendation to make builds go…	25 days ago
tests	Fix weflow test	13 days ago
udfs	Create BigQuery wrappers to avoid unnecessary repitition in common tasks	a month ago
weflow	Improvements to the BigQueryViewBuilder (#233)	4 days ago
.gitignore	Rename test. Update test to fail when a DAG can't be imported. Added …	5 months ago
README.md	Add some docs to README.md for weflow and UDFs	3 months ago
requirements.txt	Bumping Airflow to 1.7.1.2	17 days ago
run-tests.sh	Make airflow-dags PEP8 compliant, and update flake8 to enforce 120 ch…	4 months ago

DAG development

1. Install gcloud
2. Run `gcloud auth login`
3. Install/start Airflow
4. Add a Google cloud platform connection (just set project_id)

DAG development

DataInfra / airflow-dags

Watch 2 | Star 0 | Fork 5

Code | Issues 0 | Pull requests 1 | Pulse | Graphs

DI-319 Add wepay.reports to ETL pipeline #234

Merged **ChrisRiccomini** merged 1 commit into `DataInfra:master` from `ChrisRiccomini:DI-319` 4 days ago

Conversation 0 | Commits 1 | Files changed 1 | +23 −0

ChrisRiccomini commented 4 days ago

No description provided.

`DI-319 Add wepay.reports to ETL pipeline` `f8aa4f5`

ChrisRiccomini merged commit **4495036** into `DataInfra:master` 4 days ago

TeamCity commented on `f8aa4f5` 4 days ago

TeamCity DataInfra :: Airflow DAGs :: Pull requests Build 315 is now running

TeamCity replied 4 days ago

TeamCity DataInfra :: Airflow DAGs :: Pull requests Build 315 outcome was **SUCCESS**
Summary: Tests passed: 14 Build time: 00:18:15

Labels
None yet

Milestone
No milestone

Assignee
No one assigned

2 participants

DAG development

DataInfra / airflow-dags

Watch 2 | Star 0 | Fork 5

Code | Issues 0 | Pull requests 1 | Pulse | Graphs

DI-319 Add wepay.reports to ETL pipeline #234

Merged ChrisRiccomini merged 1 commit into `DataInfra:master` from `ChrisRiccomini:DI-319` 4 days ago

Conversation 0 | Commits 1 | Files changed 1 | +23 −0

ChrisRiccomini commented 4 days ago

No description provided.

`DI-319 Add wepay.reports to ETL pipeline` `f8aa4f5`

ChrisRiccomini merged commit **4495036** into `DataInfra:master` 4 days ago

TeamCity commented on `f8aa4f5` 4 days ago

TeamCity DataInfra :: Airflow DAGs :: Pull requests Build 315 is now running

TeamCity replied 4 days ago

TeamCity DataInfra :: Airflow DAGs :: Pull requests Build 315 outcome was **SUCCESS**
Summary: Tests passed: 14 Build time: 00:18:15

Labels
None yet

Milestone
No milestone

Assignee
No one assigned

2 participants

DAG testing

- flake8
- Code coverage

DAG testing

- Test that the scheduler can import the DAG without a failure
- Check that the owner of every task is a known team
- Check that the email of every task is set to a known team

DAG deployment

- CRON (ironically) that pulls from airlfow-dags every two minutes

```
$ cat ~/refresh-dags
#!/bin/bash
git -C /etc/airflow/dags/dags/dev clean -f -d
git -C /etc/airflow/dags/dags/dev pull
```

- Webserver/scheduler restarts happen manually (right now)
- DAGs toggled off by default

DAG characteristics

- (almost) All work happens off Airflow machine
- Fairly homogenous operator usage (GCP)
- Idempotent (re-run a DAG at any time)
- ETL DAGs are very small, but there are many of them

Questions?

(We're hiring)

Addendum

Usage

	On	db_monolith_wepay_batch_calls_data_quality	20 */2 * * *	di	6	
	On	db_monolith_wepay_batch_calls_monthly	0 7 2 * *	di	1	
	On	db_monolith_wepay_batch_calls_partition_15m	2,17,32,47 * * * *	di	2	
	On	db_monolith_wepay_batch_calls_partition_1d	0 4 * * *	di	2	