4 tips that help you leverage your Airflow pipelines

About me

- Aliaksandr Sheliutsin
- 6+ years in Data Engineering
- 1.7 - 2.7 Airflow experience

1

You can create typical DAGs based on a CSV file

CASE:

// Catchup pipelines for historical data download

// Triggers for support dags

name	snapshot_id	blockchain_name	dvpc_start_date	delegate_changes_start_date	delegate_histories_start_date
balancer	balancer.eth	ethereum	2020-03-17	2020-03-11	2020-05-04
ens	ens.eth	ethereum	2020-03-17	2020-03-11	2020-05-04
uniswap	uniswap	ethereum	2020-03-17	2020-03-11	2020-05-04
ampleforth	ampleforthorg.eth	ethereum	2020-03-17	2020-03-11	2020-05-04
pooltogether	pooltogether.eth	ethereum	2020-03-17	2020-03-11	2020-05-04
nexusmutual	community.nexusmutual.eth	ethereum	2020-03-17	2020-03-11	2020-05-04
hop	hop.eth	ethereum	2020-03-17	2020-03-11	2020-05-04
optimism	opcollective.eth	optimism	2022-05-26	2022-05-26	2022-04-26
1inch	1inch.eth	ethereum	2020-03-17	2020-03-11	2020-05-04
orangedao	orangedaoxyz.eth	ethereum	2020-03-17	2020-03-11	2020-05-04

PROS:

Easy to add a new entity

Less code

Flexible

Isolation/Can control start date

Also covered with data validation tests

CONS:

Advanced understanding of Airflow

Kind of similar pipelines

2 You can create typical DAGs based on a sharable Config (like Google Sheets)

CASE:

// Catchup pipelines for historical GA pipelines

PROS:

Other people can create Airflow DAGs

Changes are available without deployments

CONS:

Performance

You need to handle user experience

// Documentation
// Input data validation
// Error Notifications

3 Autogenerate DAGs based on the Airflow Entities (Connection, Variables)

CASE:

☐	✎ 🗑	processing_neo	mysql	Historical data from processing
☐	✎ 🗑	processing_trinity	mysql	Historical data from processing

PROS:

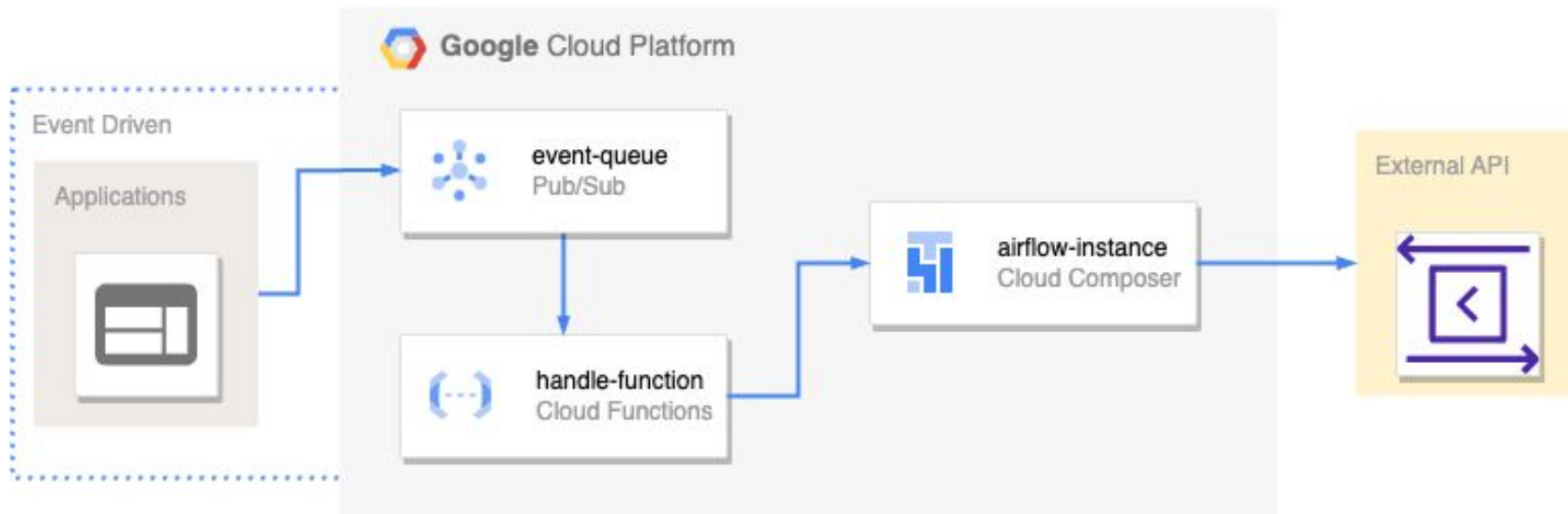No need to handle External API calls

CONS:

Configurations can be lost

Higher load on Metadata DB

4

Make your Airflow pipeline event-driven

CASE:

PROS:

Results of execution do not rely on a schedule

Don't have tons of one-minute Dag Runs

CONS:

This way will not work with a high volume of events

Have questions?

// **Linkedin:** https://www.linkedin.com/in/aliaksandr-sheliutsin/

// **Telegram:** @asheliutsin