DAG Authoring without PhD

Filip Knapik, Rafał Biegacz
About us

Filip Knapik
Group Product Manager
Cloud Composer

Working with Airflow for ~4 years
Ex-Product Manager for Google Workflows

Rafal Biegacz
Senior Eng Manager
Cloud Composer

Working with Airflow for ~4 years
Member of Airflow Summit Organizing Team
Contacting Us / Where to Find Us

During the conference visit us at Google Booth to learn more about:
- Serverless Composer
- Disaster Recovery & Data Lineage support in Composer
- Support for Public Sector and Assured Workloads

Fill in the form: [bit.ly/airflow-summit-2023-composer](bit.ly/airflow-summit-2023-composer) if you would like to meet, request more information or you are interested in getting a voucher for GCP credits.
DAG Authoring is non-trivial

Business Need

DAG Development

- APIs
- 100x
- API Docs
- Operators
- Airflow
- Python
- IDE + CI/CD

DAG
Can it be any easier?

LLM-based Generative AI model acting as a translation layer

I need a DAG that ...
How to get there?

AI model preparation

Prompts + DAGs

Fine-tuning

New LLM Model

Predictions

Existing LLM model

Business Need

DAG

Prompt

DAG that runs a query in BigQuery at 8 AM every day

Airflow DAG running 10 bash commands in parallel

...
The challenge

DAGs + Prompts

Training

The LLM Model

Predictions

Business Need

DAG

The hardest part!
Where we are

Prompt: Write Airflow DAG that runs BigQuery query using BigQueryInsertJobOperator to get number of GitHub commits in May 2022. Please, use public dataset for github data published by Google Cloud Platform.
Our proposal

Establish an open source Airflow DAGs + Prompts repository to enable LLM models creation & experimentation by the community

- Google establishes open source repository
- Google donates its Airflow DAGs training data
- Community joins in and contributes with its own DAGs examples
- Community can build its own DAG code generation models and tools
Contribution process

1. Clone

2. Add DAGs & prompts

3. Presubmit checks
   - Python correctness
   - Registration in DAG Bag
   - Prompts quality

4. Merge

Your Repository → Pull Request → Main Repository
Contribution guidelines

- The more (DAG, prompt) pairs are contributed, the better
- Ensure that DAGs and prompts are:
  - Free of any PII references
  - Following latest Airflow development practices
  - Using particular versions of python modules and provider packages
  - As diverse as possible
- Use examples that you can contribute under Apache 2.0 license
How to use it?

1. Clone Main Repository
2. Run a data preparation script (may exist for your AI stack)
3. Train your model

Your model
Create a tuned model

- Tuning type
- Model details
- Tuning dataset

4 Evaluation (optional)

Model evaluation

Model evaluation creates a batch response using the evaluation dataset to generate new statistics.

Enable model evaluation

Dataset location *

 gs://cloudfunimagebucket

The Cloud Storage location where the JSONL file will be stored.

CONTINUE
Next Steps

1. Composer team will send a call for action email on Airflow Dev list
2. All community members are encouraged to join us!
Questions?

Optionally share some contact info like email, blog or social media handles
DAG Authoring without PhD

Filip Knapik, Rafał Biegacz