

Airflow Pandawans



Elementos que são utilizados em uma stack de dados.



QUEM SOU EU?

Arthur Gonzaga (Tutis)

26 Anos

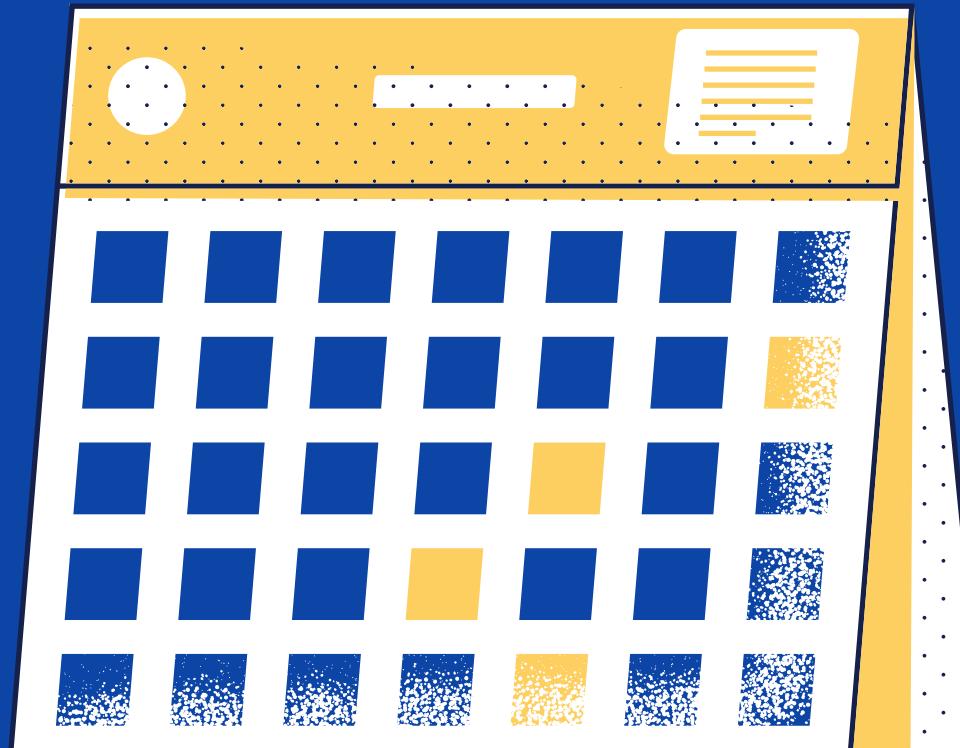
Analytics Engineer no QuintoAndar

Engenharia Eletrônica na UnB - FGA

Futebol, cerveja e pagode



O QUE É O AIRFLOW?



Ferramenta utilizada para orquestração dos pipelines e tarefas, geralmente é bem usado na área de dados. É escrita em python e pertence ao guarda-chuva da Apache.

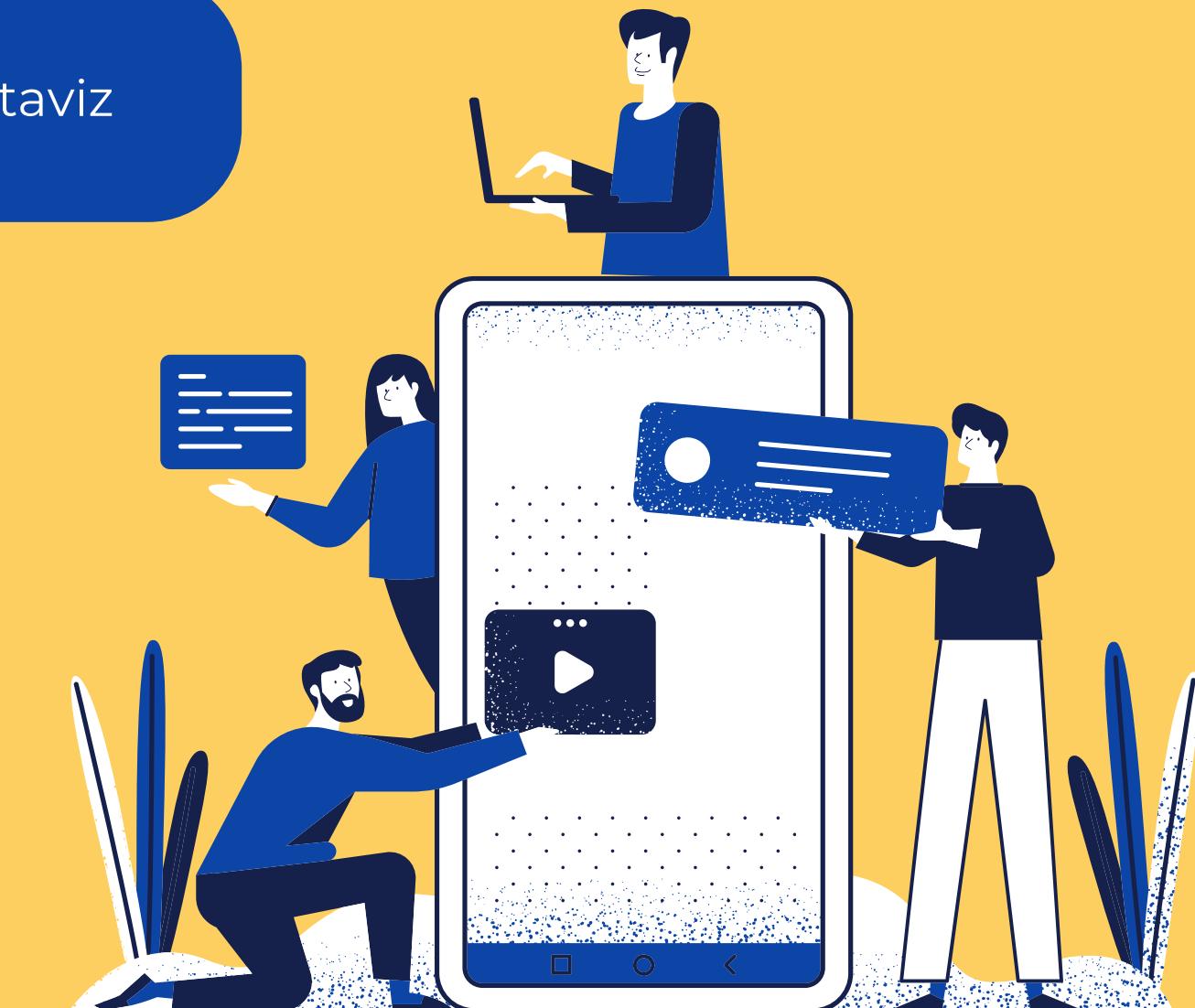
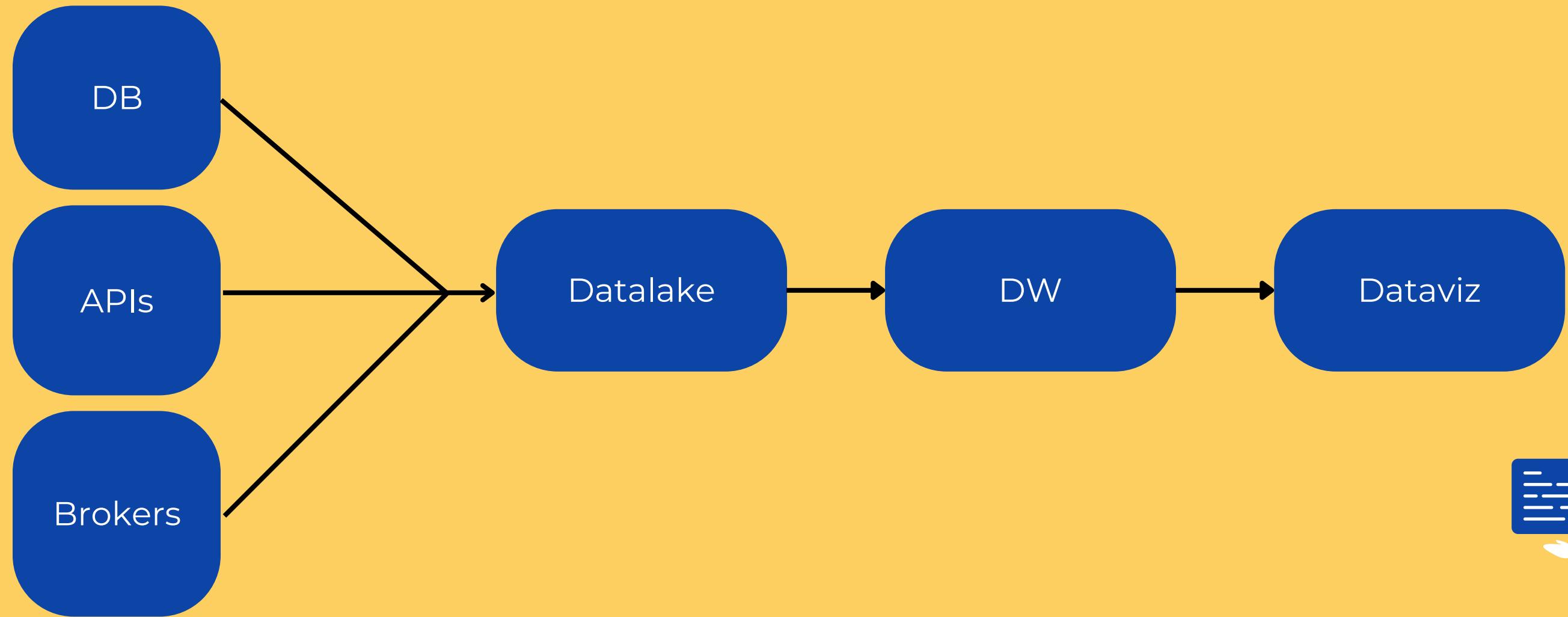
Como posso executar o Airflow?

- Disponível nas clouds (MWAA e Composer)
- Disponível em docker para rodar na máquina local
- Disponível também como instalável para execução em ambiente python.

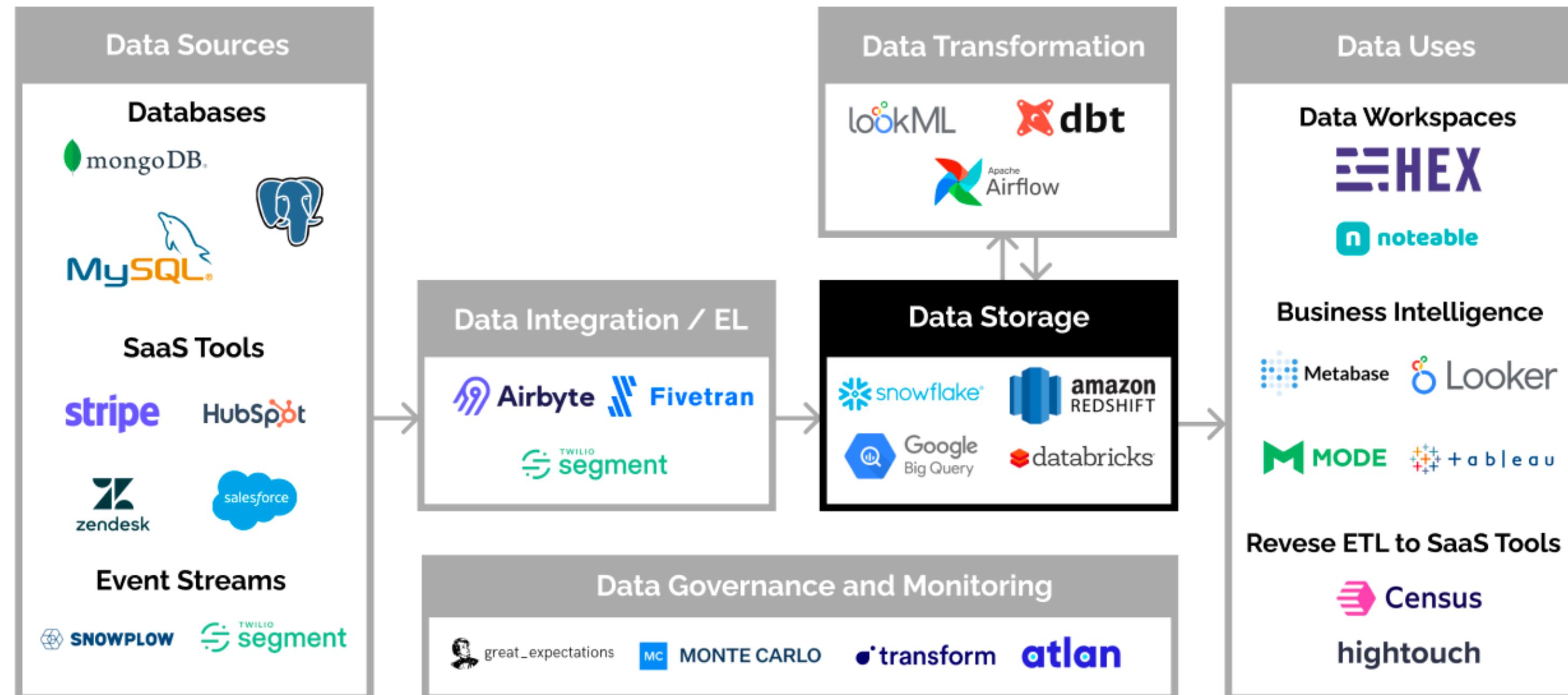
Quem usa?

- QuintoAndar
- XP Inc
- Airbnb

Modern Data Stack

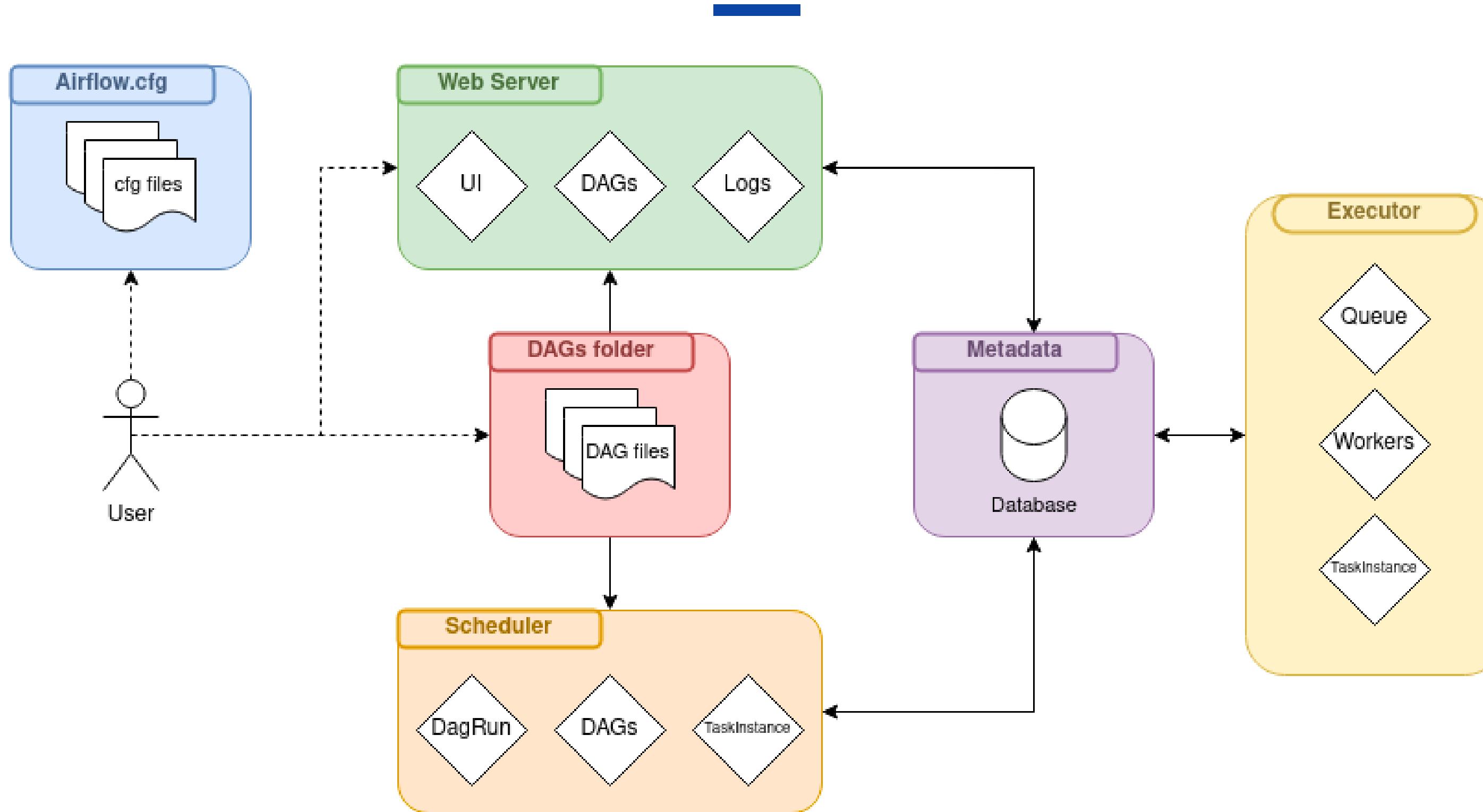


Modern Data Stack

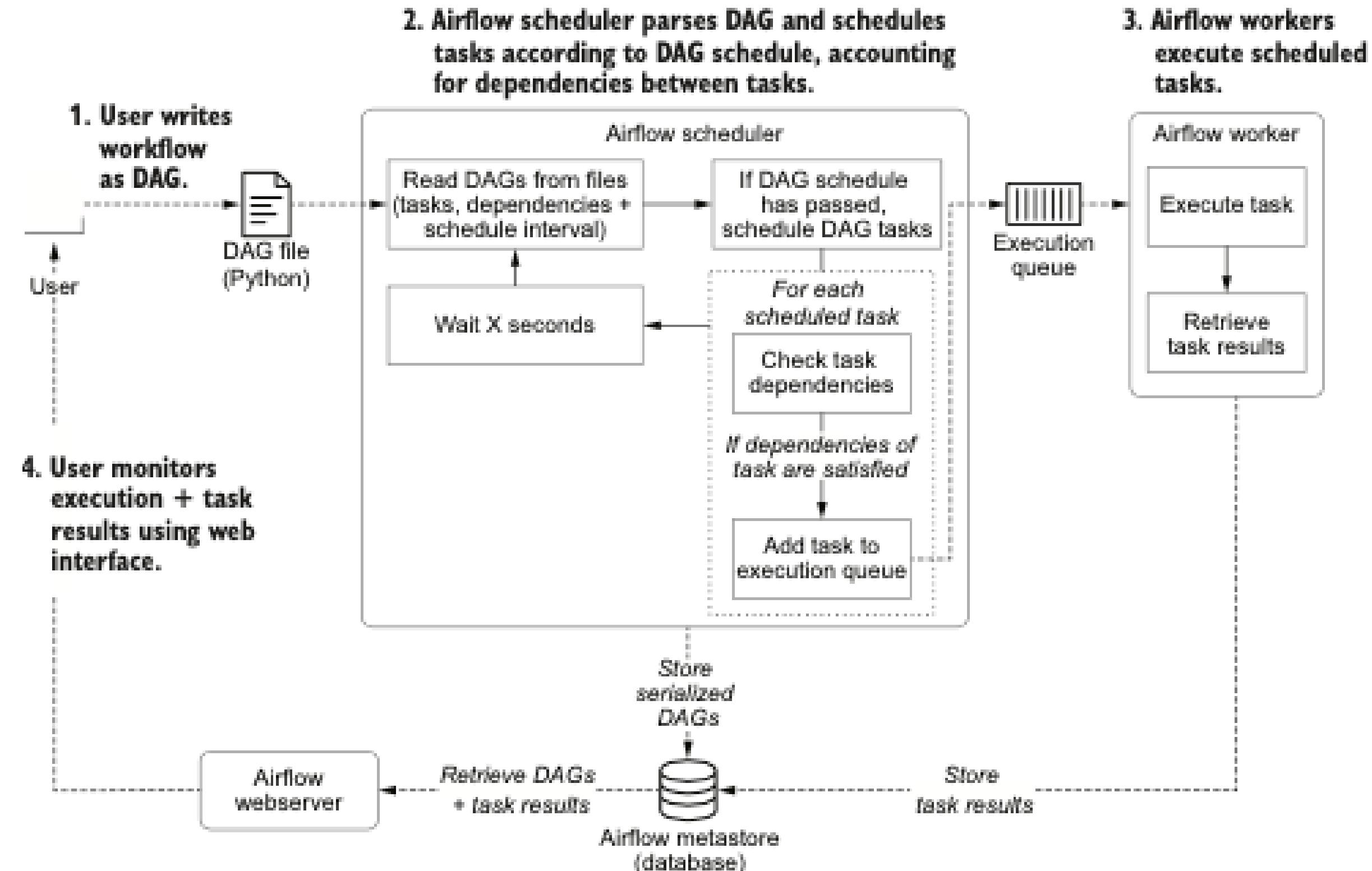


@tanayj

ARQUITETURA DO AIRFLOW



ARQUITETURA DO AIRFLOW



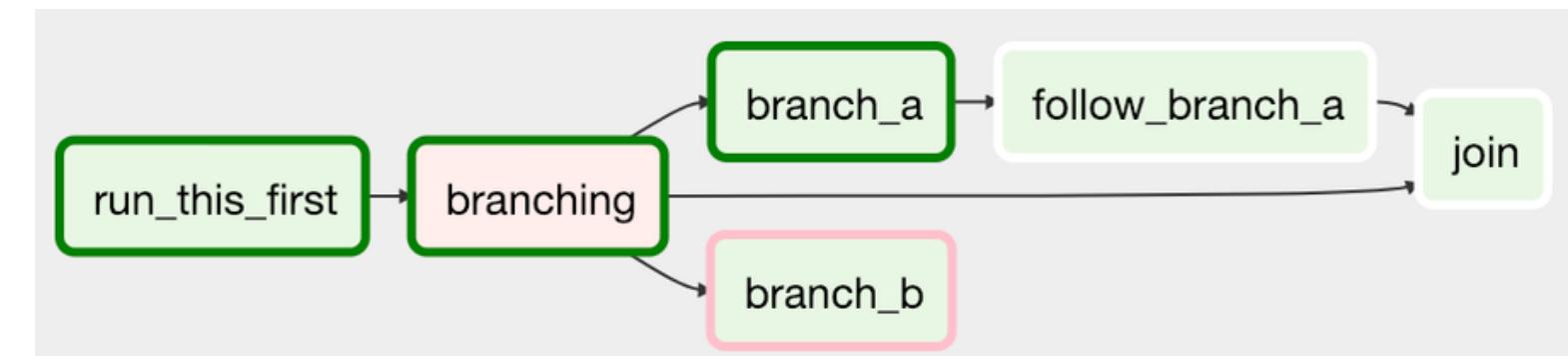


Hora de
vermos a
carinha do
Airflow!

DAGS

DAG (Directed Acyclic Graph):

- É um grafo que permite organizar de certa maneira o pipeline de dados, separando o passo a passo de execução e priorizando tarefas de acordo com as regras pré-definidas.



ETL / ELT



O que é o processo de ETL ou ELT?

São as etapas de ingestão do dado, onde você pode realizar operações numéricas, enriquecimento com regras de negócio ou particionamento por data. Exemplos:

Extração: requisição de uma API

Transformação: cast de uma variável

Carregamento: salvar em arquivo

RAW

date	client	cel
2022/01/01	sergi dest	+556199123123
2022-01-03	gareth BALE	666199321312
2022.01.16	weah	99999999

CLEAN

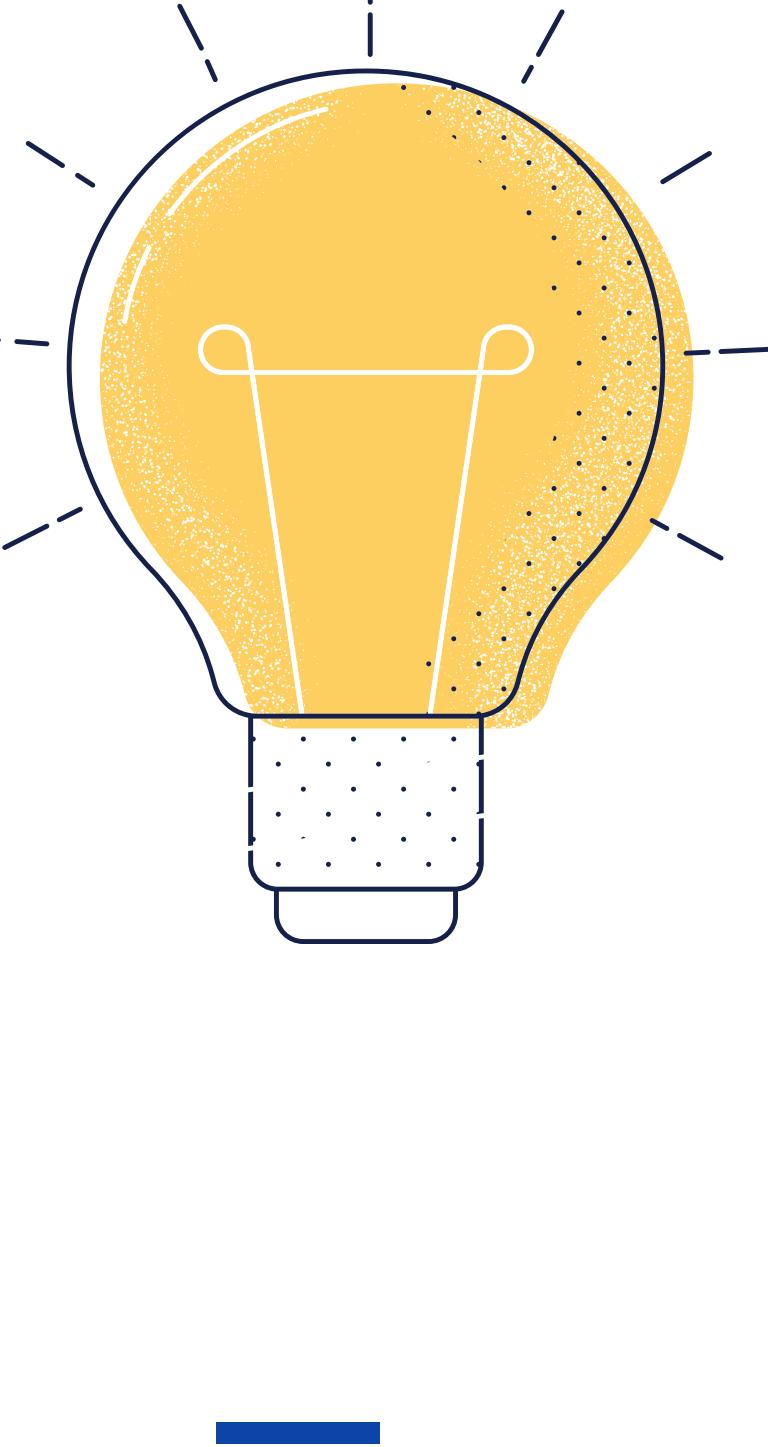
date	client	country	region	cel
2022-01-01	Dest	55	61	99123123
2022-01-03	Bale	66	61	99321312
2022-01-16	Weah	NULL	NULL	99999999

CURATED

month	clients	most_common_region
jan	3	brasilia

Bora executar umas DAGs?





Livro referência: Data
Pipelines with Apache
Airflow

