Airflow

João Pedro V. Pinheiro

joaopedro.pinheiro88@gmail.com

29-Abr-2022

Airflow

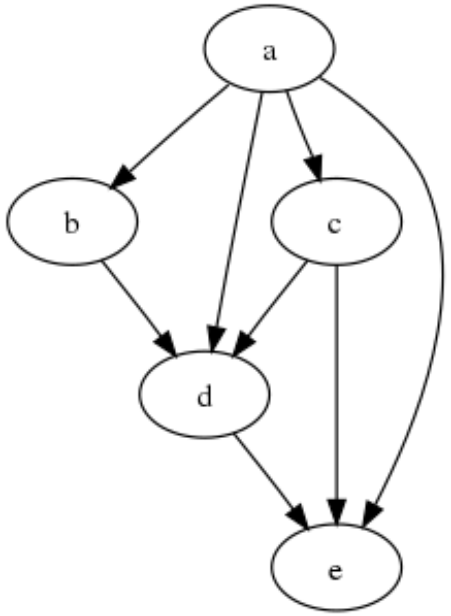
Definição e Origem

- Uma plataforma para criar, agendar, monitorar e orquestrar *workflows*
- O projeto foi criado pela empresa *Airbnb* em 2014, mas já foi concebido com o conceito *open source*
- O projeto foi adotado pela *Apache Software Foundation* em 2019
- A linguagem base escolhida para o projeto foi *Python*

Airflow

Conceitos Básicos

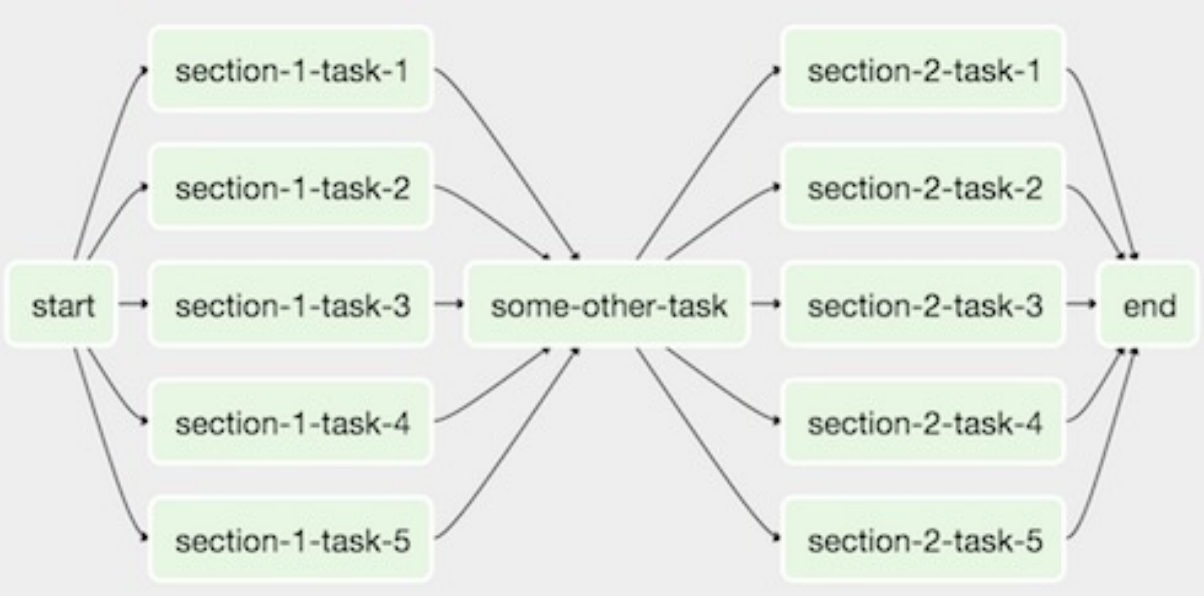
- *Directed Acyclic Graph* (DAG):
  - Um grafo acíclico direcionado (DAG) é um conjunto de vértices e arestas no qual as arestas direcionam um vértice ao outro, em que não haja formação de um laço fechado
  - Dessa forma, loops não são permitidos dentro dos *workflows*



Airflow

Conceitos Básicos

- *Task*:
  - É a unidade básica de computação do Airflow
  - São os vértices das DAGs e as dependências entre elas conectam esses vértices, formando um grafo



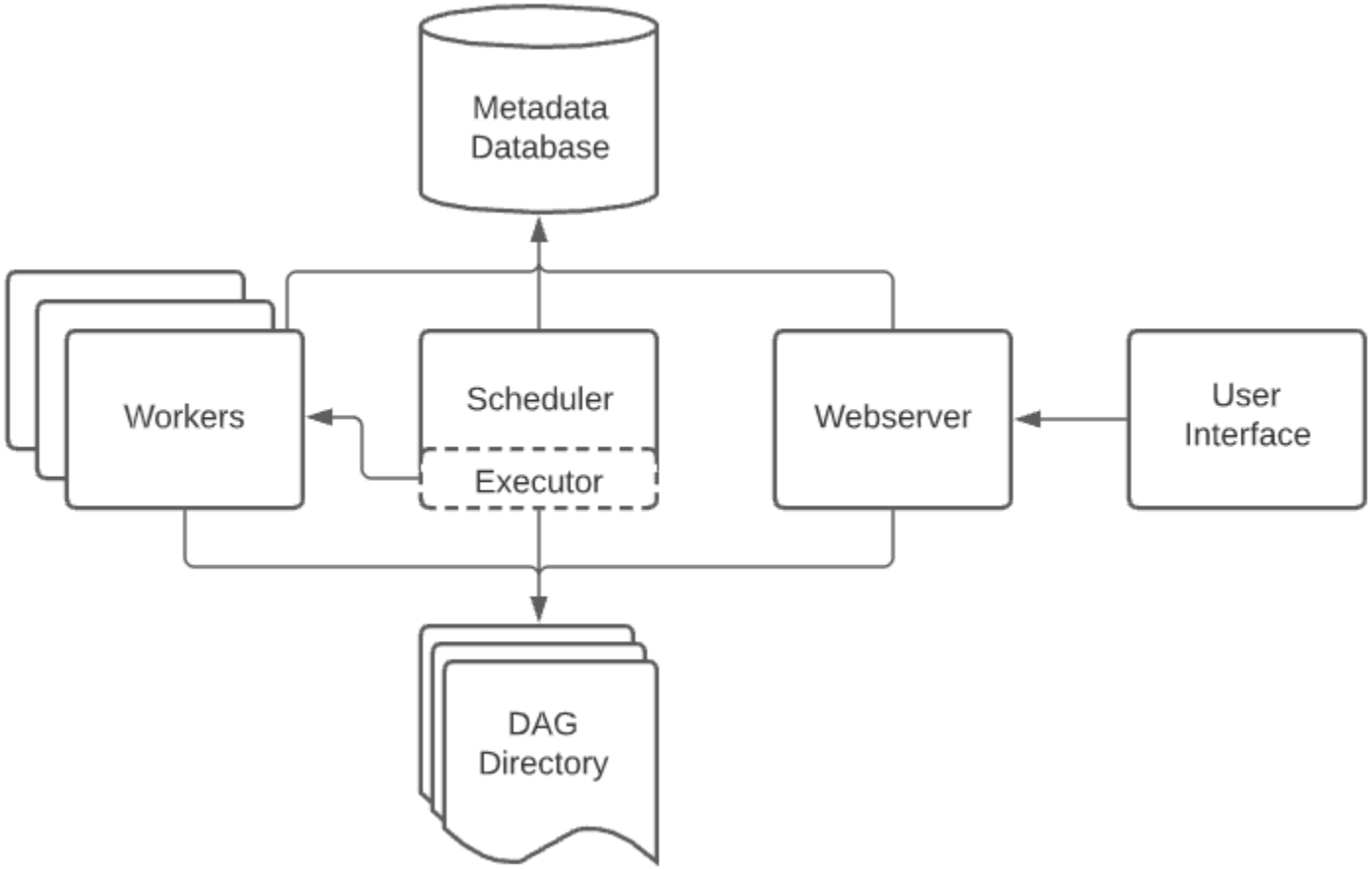
Airflow

Conceitos Básicos

- Existem três tipos básicos de *Task*:
  - Operadores (*Operators*)
    - modelos de tarefas predefinidos que você pode agrupar rapidamente para criar a maior parte das suas DAGs
    - vasta quantidade de operadores existentes gerenciados pela comunidade
  - Sensores (*Hooks*)
    - uma subclasse especial de Operadores que dependem que um evento externo aconteça
    - na demonstração, utilizamos *Hooks* para interação com o banco de dados
  - Funções genéricas
    - uma função qualquer em Python decorada com @task
    - na demonstração, utilizamos dessa forma para instanciar os Hooks

Airflow

Arquitetura



Airflow

Arquitetura

- Arquitetura composta pelos seguintes componentes:
  - Agendador (*Scheduler*)
    - lida tanto com a questão do disparo dos *pipelines* (*workflows*), quanto o envio das *tasks* para execução do Executor
  - Executor
    - lida com as *taks* em execução
    - utiliza *workers* para melhor controle e paralelização das execuções
  - Servidor Web
    - interface gráfica para interação dos usuários
    - utilizada para inspeção, disparo manual das DAGs, controle de variáveis, conexões, ...

Airflow

Arquitetura

- Continuação dos componentes da arquitetura:
  - Pasta de DAGs
    - arquivos Python com a definição das tarefas
    - consumido pelo *Scheduler* e Executor
  - Banco de Dados para Metadados
    - utilizado pelo *Scheduler*, Executor e Servidor Web para persistir o estado da aplicação e de seus componentes

Obrigado!

João Pedro V. Pinheiro

joaopedro.pinheiro88@gmail.com

29-Abr-2022