Hands On

Apache Airflow

Introdução

O \*\*Apache Airflow\*\* é uma plataforma de código aberto usada para \*\*orquestração de workflows e pipelines de dados\*\*. Ele permite que você crie, agende e monitore fluxos de trabalho de forma programática, utilizando o Python como linguagem principal.

### Componentes principais:

1. \*\*DAGs (Directed Acyclic Graphs):\*\* Representam o fluxo de tarefas que precisam ser executadas em uma sequência específica, sem ciclos.

2. \*\*Operadores:\*\* Componentes que definem as tarefas individuais, como execução de scripts, movimentação de dados ou consultas SQL.

3. \*\*Scheduler:\*\* Responsável por agendar a execução das tarefas com base na definição das DAGs.

4. \*\*Executor:\*\* Gerencia a execução das tarefas (por exemplo, LocalExecutor, CeleryExecutor).

5. \*\*Web Interface:\*\* Uma interface gráfica que permite visualizar, monitorar e gerenciar as DAGs.

### Benefícios:

- \*\*Escalabilidade:\*\* Pode ser usado em pequenos projetos ou em arquiteturas complexas.

- \*\*Flexibilidade:\*\* Oferece integração com diversas ferramentas (bancos de dados, APIs, Hadoop, etc.).

- \*\*Automatização:\*\* Facilita a execução recorrente e programada de tarefas.

Airflow é amplamente usado em \*\*pipelines de ETL\*\*, integração de dados e workflows de machine learning. É ideal para cenários onde tarefas precisam ser altamente organizadas e monitoradas.