



EDUARDO QUETZAL DELGADO PIMENTEL

Materia: Computación Tolerante a Fallas

Fecha:15/04/2024

Codigo:217239716

## Introducción a Apache Airflow: Gestión de Workflows y Orquestación de Tareas

Apache Airflow es una plataforma de código abierto utilizada para programar, supervisar y gestionar flujos de trabajo (workflows) de manera dinámica. Desde su introducción en 2015 por Airbnb, ha ganado popularidad en la comunidad de desarrollo y en empresas de diversos sectores debido a su flexibilidad, escalabilidad y capacidades de orquestación de tareas.

### ¿Qué es Airflow?

Airflow permite a los equipos definir, programar y ejecutar flujos de trabajo complejos con facilidad. Su arquitectura se basa en conceptos clave como DAGs (Directed Acyclic Graphs), que representan los flujos de trabajo como grafos dirigidos sin ciclos, lo que facilita la visualización y comprensión de las dependencias entre las tareas.

### Características Principales:

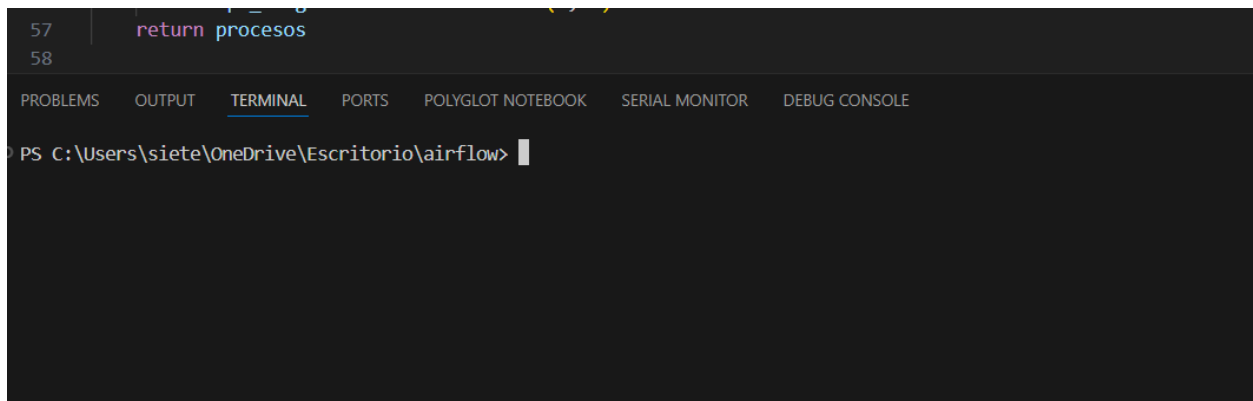
1. **Programación Declarativa:** Airflow utiliza Python para definir flujos de trabajo de forma declarativa, lo que significa que los desarrolladores pueden expresar las dependencias y la lógica de sus tareas de manera clara y legible.
2. **Escalabilidad:** Puede manejar flujos de trabajo de cualquier tamaño, desde simples hasta complejos, escalando horizontalmente para satisfacer las demandas de procesamiento.
3. **Monitoreo y Gestión:** Proporciona una interfaz de usuario intuitiva para monitorear el estado de los flujos de trabajo, ver el historial de ejecuciones y gestionar los recursos de manera eficiente.
4. **Extensibilidad:** Airflow es altamente personalizable y extensible, lo que permite a los equipos integrar fácilmente nuevas fuentes de datos, operadores y conectores con su infraestructura existente.

### Casos de Uso:

1. **ETL (Extract, Transform, Load):** Automatización de procesos de extracción, transformación y carga de datos entre sistemas heterogéneos.
2. **Programación de Tareas:** Ejecución programada de tareas como procesamiento de datos, generación de informes o notificaciones.
3. **Machine Learning Pipelines:** Orquestación de flujos de trabajo para entrenar modelos, realizar inferencias y gestionar el ciclo de vida de los modelos de machine learning.
4. **Automatización de Infraestructura:** Gestión de despliegues, configuración y monitoreo de infraestructura a través de flujos de trabajo definidos por código.

## Desarrollo:

Aquí se abre el entorno de Python, pero como no pude correr el airflow ya que me daba error al momento de iniciar el db, me decía que estaba mal configurado.



The screenshot shows a code editor with a dark theme. At the top, there are tabs for 'PROBLEMS', 'OUTPUT', 'TERMINAL', 'PORTS', 'POLYGLOT NOTEBOOK', 'SERIAL MONITOR', and 'DEBUG CONSOLE'. The 'TERMINAL' tab is active, showing a command prompt with the path 'PS C:\Users\siete\OneDrive\Escritorio\airflow>'. Above the terminal, there is a snippet of Python code with line numbers 57 and 58. Line 57 contains the text 'return procesos'.

Así que mejor lo pase a Linux y así es como se ve la instalación de todas las dependencias

```
aiosignal-1.3.1 alembic-1.11.1 anyio-3.7.1 apache-airflow-2.6.3 apache-airflow-providers-common-sql-1.5.2 apache-airflow-providers-ftp-3.4.2 apache-airflow-providers-http-4.4.2 apache-airflow-providers-imap-3.2.2 apache-airflow-providers-sqlite-3.4.2 apispec-5.2.2 argcomplete-3.1.1 asgiref-3.7.2 async-timeout-4.0.2 attrs-23.1.0 blinker-1.6.2 cachelib-0.9.0 cattrs-23.1.2 certifi-2023.5.7 cffi-1.15.1 charset-normalizer-3.1.0 click-8.1.4 clickclick-20.10.2 colorama-0.4.6 connexion-2.14.2 cron-descriptor-1.4.0 croniter-1.4.1 cryptography-3.4.8 dill-0.3.1.1 dnspython-2.3.0 docutils-0.20.1 email-validator-1.3.1 exceptiongroup-1.1.2 frozenlist-1.3.3 google-re2-1.0 graphviz-0.20.1 greenlet-2.0.2 h11-0.14.0 httpcore-0.16.3 httpx-0.23.3 idna-3.4 importlib-metadata-4.13.0 importlib-resources-5.12.0 inflection-0.5.1 itsdangerous-2.1.2 jsonschema-4.18.0 jsonschema-specifications-2023.6.1 lazy-object-proxy-1.9.0 limits-3.10.1 linkify-it-py-2.0.2 lockfile-0.12.2 markdown-it-py-3.0.0 marshmallow-3.19.0 marshmallow-enum-1.5.1 marshmallow-oneofschema-3.0.1 marshmallow-sqlalchemy-0.26.1 marshmallow-plugin-0.4.0 mdurl-0.1.2 multidict-6.0.4 ordered-set-4.1.0 packaging-21.3 pathspec-0.9.0 pendulum-2.1.2 pkgutil-resolve-name-1.3.10 pluggy-1.2.0 pri-1.2.1 psutil-5.9.5 pycparser-2.21 pydantic-1.10.11 pyparsing-3.1.0 python-dateutil-2.8.1 python-dateutil-2.8.2 python-nvd3-0.15.0 python-slugify-8.0.1 pytz-2023.3 pytzdata-2020.1 referencing-0.29.1 requests-2.31.0 requests-toolbelt-1.0.0 rfc3986-validator-0.1.4 rfc3986-1.5.0 rich-13.4.2 rich-argparse-1.2.0 rpds-py-0.8.4 setproctitle-1.3.2 six-1.16.0 sniffio-1.3.0 sqlparse-0.4.4 tabulate-0.9.0 termcolor-1.1.0 termcolor-1.1.0 text-unidecode-1.3 typing-extensions-4.7.1 uc-micro-py-0.9.7 unidecode-1.1.4 urllib3-1.26.16 uvicorn-0.15.0 uvloop-0.15.0
```

Entonces después de instalar el airflow hice un script en Python.

```
1 from airflow import DAG
2 from airflow.operators.python_operator import PythonOperator
3 from datetime import datetime
4
5 def print_hello():
6     return 'Hola, ¡este es un mensaje impreso desde Airflow!'
7
8 # Definir los argumentos de la DAG
9 default_args = {
10     'owner': 'usuario',
11     'depends_on_past': False,
12     'start_date': datetime(2024, 4, 14),
13     'email_on_failure': False,
14     'email_on_retry': False,
15     'retries': 1,
16     'retry_delay': timedelta(minutes=5),
17 }
18
19 # Definir la DAG
20 dag = DAG(
21     'mi_primer_flujo_de_trabajo',
22     default_args=default_args,
23     description='Un simple flujo de trabajo con Airflow',
24     schedule_interval=timedelta(days=1),
25 )
26
27 # Definir las tareas
28 hello_operator = PythonOperator(
29     task_id='print_hello',
30     python_callable=print_hello,
31     dag=dag,
32 )
33
34 # Definir la secuencia de tareas
```

### Conclusión:

Apache Airflow es una herramienta poderosa para la orquestación de flujos de trabajo que ofrece a los equipos la capacidad de automatizar y gestionar procesos complejos de manera eficiente. Su flexibilidad, escalabilidad y amplia comunidad de usuarios y contribuyentes hacen de Airflow una opción atractiva para empresas que buscan mejorar la eficiencia operativa y la confiabilidad de sus sistemas.

