Reponse

1. Quelle serait votre stratégie de mise en production de votre ETL ?

Pour mettre en production mon processus ETL, je ferais les choses suivantes :

- **Containerisation avec Docker** : Je créerais une image Docker pour encapsuler l'ensemble de l'application ETL, incluant toutes les dépendances nécessaires comme les bibliothèques Python pour l'ETL, Airflo, ...
- **Utilisation d'Apache Airflow** : J'utiliserais Apache Airflow pour orchestrer le processus ETL. Cela permet de gérer les tâches de manière modulaire, de suivre les dépendances entre les différentes étapes et d'assurer une exécution robuste.
- **Tests unitaires et d'intégration** : Avant la mise en production, je mettrais en place des tests unitaires et d'intégration pour m'assurer que chaque composant fonctionne comme prévu.
- **Surveillance et journalisation** : Je mettrais en place des mécanismes de surveillance pour suivre l'exécution des tâches ETL et détecter les éventuelles erreurs. Airflow propose des fonctionnalités de journalisation qui peuvent être exploitées pour surveiller l'état des tâches.

2. Quelle serait votre stratégie pour déclencher votre ETL automatiquement chaque heure ?

Pour déclencher automatiquement mon processus ETL chaque heure, je mettrais en place les étapes suivantes :

- **Planification dans Apache Airflow** : Je configurerais un **DAG (Directed Acyclic Graph)** dans Airflow, qui représente l'ensemble du processus ETL. Dans ce DAG, j'utiliserais le paramètre schedule_interval pour définir la fréquence d'exécution. Dans notre cas pour exécuter l'ETL toutes les heures, je définirais le schedule_interval comme suit :

```python
dag = DAG(
    'my_etl_dag',
    default_args={
        'owner': 'Yann',
        'email':'amanikoney@gmail.com',
        'depends_on_past': False,
        'start_date': datetime(2024, 10, 16),
        'retries': 1,
        'retry_delay': timedelta(minutes=5),
    },
    schedule_interval='@hourly',  # Exécute toutes les heures
)
```

- **Surveillance des échecs** : Je mettrais en place des alertes pour m'informer en cas d'échec d'une exécution, afin de pouvoir intervenir rapidement en cas de problème.
- **Gestion des dépendances** : En utilisant Airflow, je peux m'assurer que les tâches se déclenchent dans l'ordre correct et que les dépendances sont respectées, ce qui permet une exécution fluide et coordonnée du processus ETL.