MISE EN PLACE D'UN PIPELINE D'ORCHESTRATION DE FLUX DE DONNÉES

Contexte et problématique

- **Situation actuelle :**
  - manuellement chaque mois :

    Nettoyage de 3 fichiers Excel (ERP, LIAISON, WEB)

    Fusion des données

    Calcul du chiffre d'affaires Identification des vins premium (z-score)

- **Problème :**
  - Processus long et répétitif
  - Risque d'erreurs humaines
  - Pas de traçabilité

- **Objectif :**
  - Automatiser le pipeline avec Kestra pour une exécution mensuelle automatique

Méthodologie

Jupyter Notebook

- Analyse données
- Identification relations entre les fichiers
- Choix Z-score

EXPLORATION (EDA)

SCRIPTS

SQL + Python

- Nettoyage
- Fusion
- Calculs
- Tests

ORCHESTRATION

Kestra

- Workflow YAML
- Docker
- Tests automatisés

Sources de données et relations

ERP.xlsx

825 lignes

LIAISON.xlsx

825 lignes

WEB.xlsx

1513 lignes

714 lignes finales

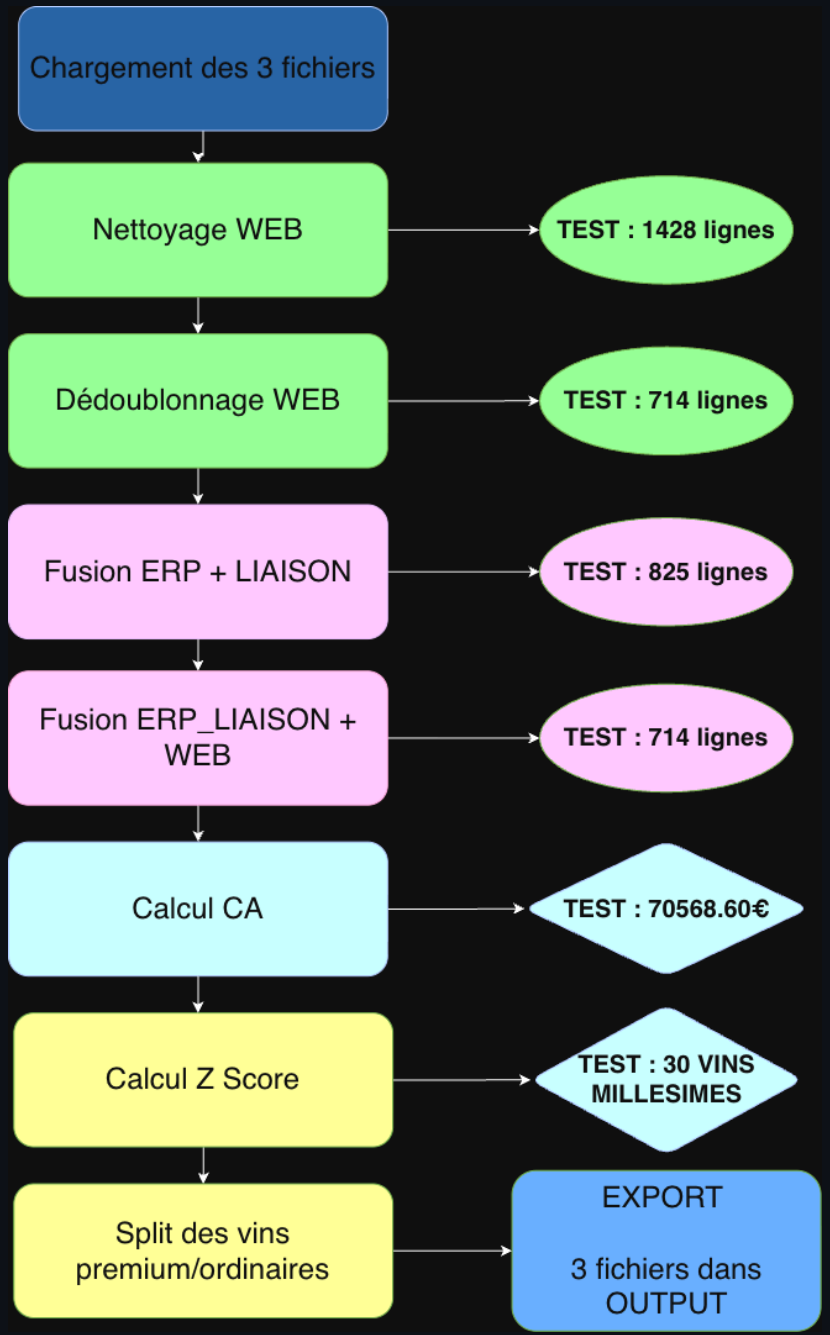
FUSION

2 INNER JOIN

ERP ↔ LIAISON :
product_id

LIAISON ↔ WEB :
id_web = sku

Flux de données - Vue d'ensemble



Stack technique

🐳 DOCKER

- Conteneurisation
- Reproductibilité
- Image personnalisée

🐍 PYTHON

- Manipulation données
- pandas (ex : lecture fichiers Excel fournis)
- scipy (z-score)

🦆 DUCKDB

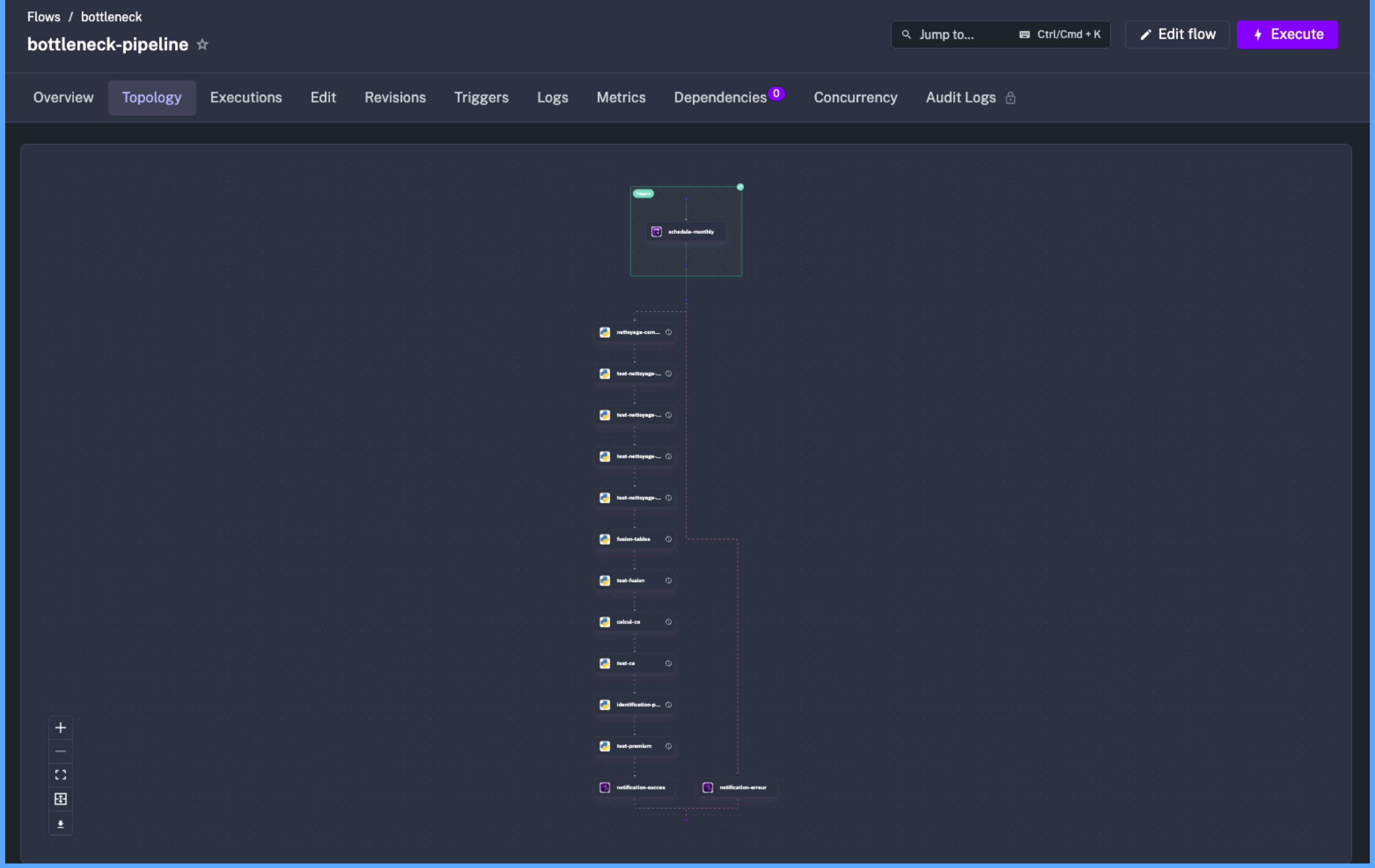
- Requêtes SQL
- Performance analytique
- Export Parquet

📊 FORMATS

Parquet (échange données)

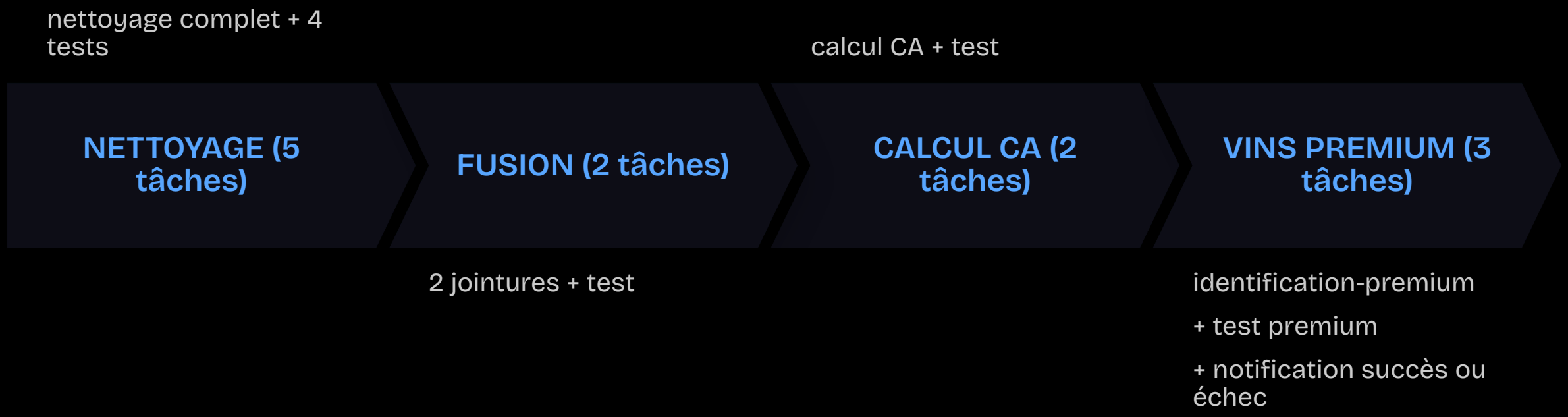
Excel (rapport CA)

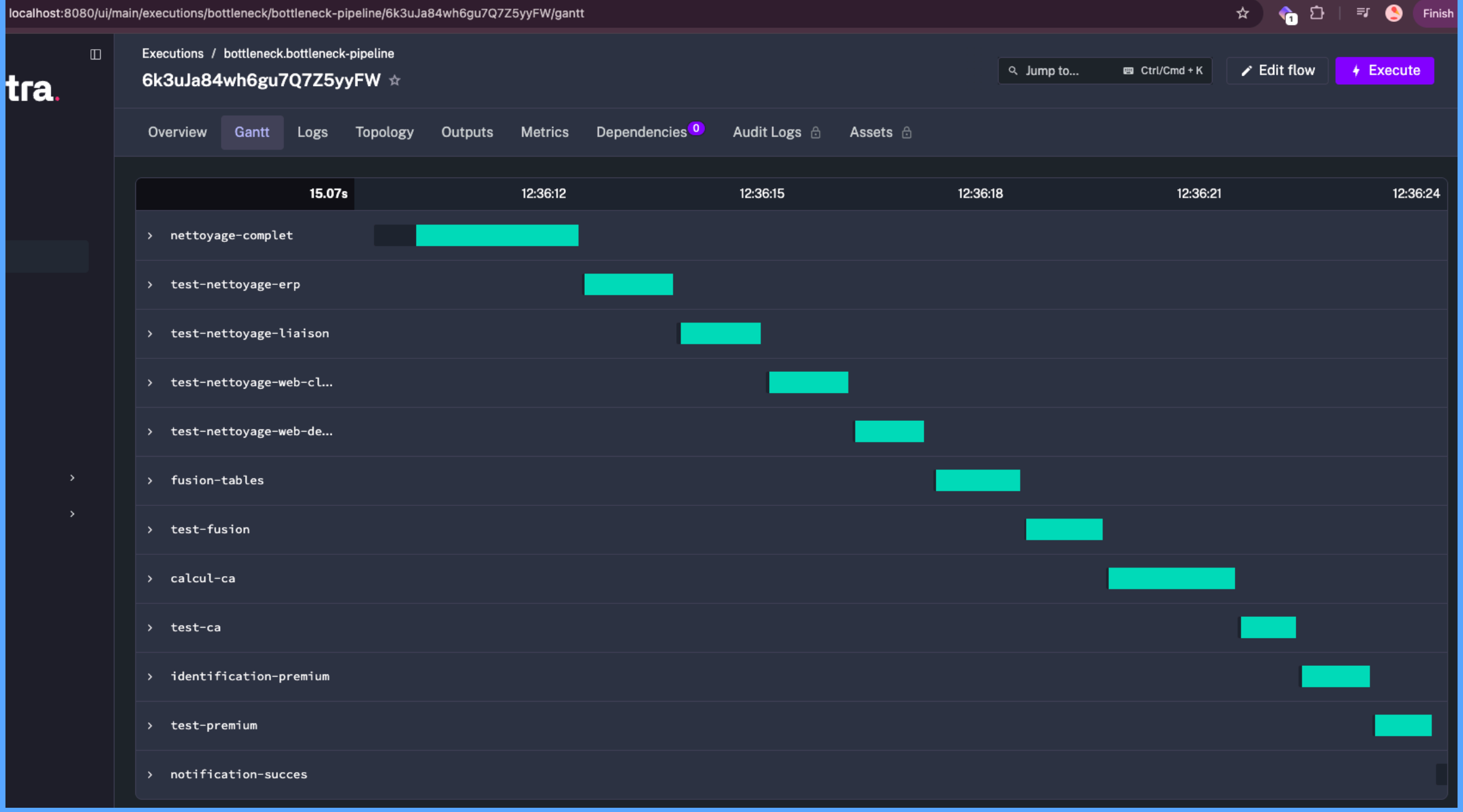
CSV (vins premium/ordinaires)



Pipeline d'orchestration (DAG) sur Kestra

Pipeline d'orchestration (DAG)





Vue Gantt du pipeline d'orchestration

7 tests pour des alertes granulaires

Point clé : Tests séparés = Identification précise des erreurs

Test	Validation	Valeur attendue	Alerte si échec
test-nettoyage-erp	Absence doublons ERP	825 lignes	❌ ERP: X lignes au lieu de 825
test-nettoyage-liaison	Absence doublons LIAISON	825 lignes	❌ LIAISON: X lignes au lieu de 825
test-nettoyage-web-clean	Suppression valeurs manquantes	1428 lignes	❌ WEB: X lignes au lieu de 1428
test-nettoyage-web-dedup	Dédoublonnage	714 lignes	❌ WEB DEDUP: X lignes au lieu de 714
test-fusion	Cohérence volumétrie jointures	714 lignes	❌ FUSION: X lignes au lieu de 714
test-ca	Cohérence CA calculé	70 568,60€	❌ CA: X€ au lieu de 70 568,60€
test-premium	Cohérence z-score > 2	30 vins	❌ PREMIUM: X vins au lieu de 30

Persistance via fichiers Parquet

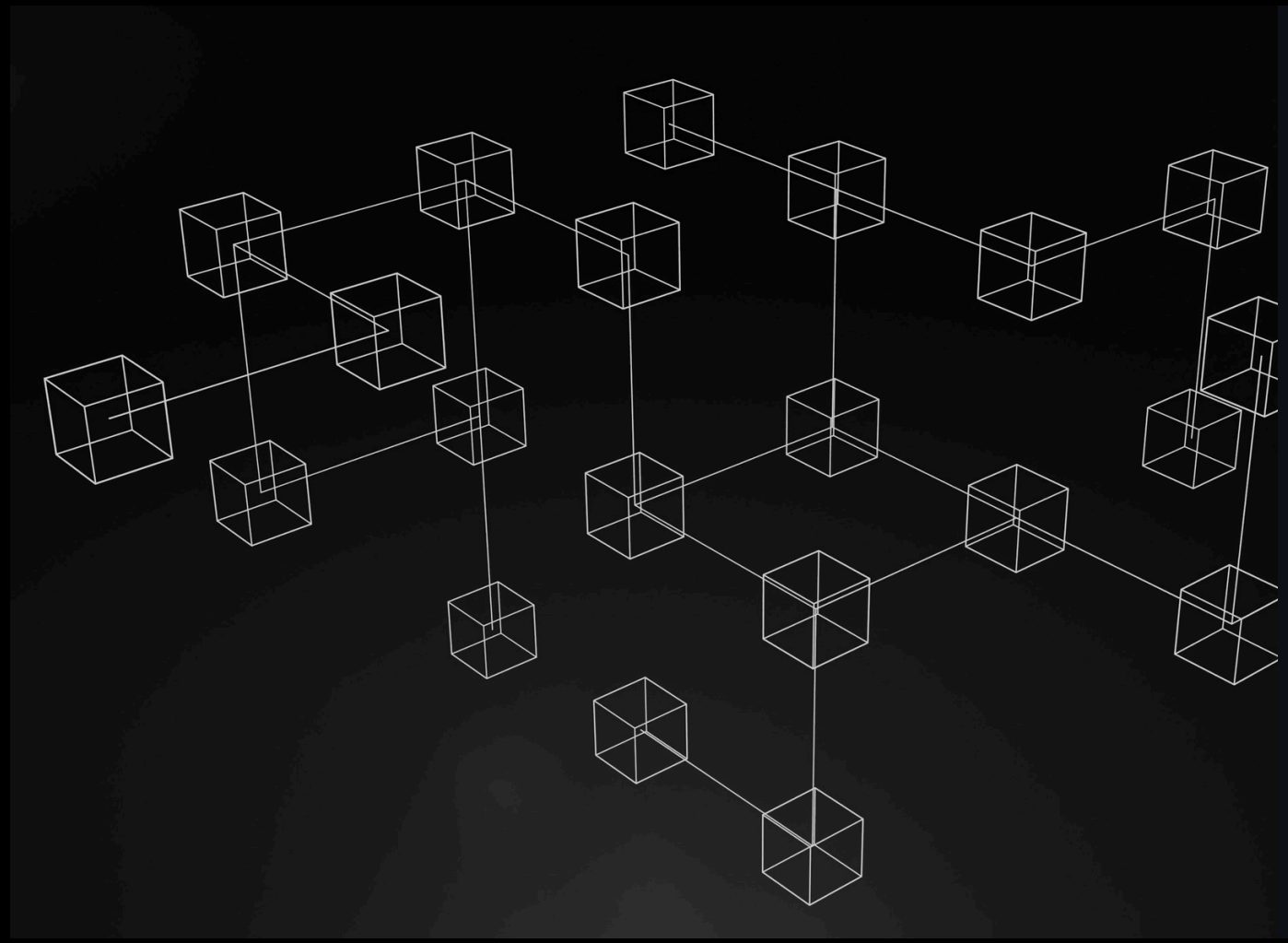
Solution :

- Export Parquet à la fin de chaque tâche
- Import Parquet au début de la tâche suivante
- Persistance des données entre containers

Problème :

- Chaque tâche = Nouveau container →
- Mémoire effacée après exécution →
- Tables DuckDB perdues

Démonstration



Prise en main de Kestra

Évolutions possibles

1

NOTIFICATIONS AUTOMATIQUES

Kestra Enterprise ou Resend API

2

ROBUSTESSE

Retry automatique (3 tentatives)

3

ARCHIVAGE

Conservation historique mensuelle des données

Comparaison mois par mois

Objectifs atteints :

Pipeline 100% automatisé

Exécution mensuelle planifiée

Tests automatisés à chaque étape

Alertes granulaires en cas d'erreur

Fichiers générés conformes aux attentes

Merci pour votre attention ! Questions ?