

A decorative graphic on the left side of the slide, consisting of a network of light blue lines and small circles, resembling a circuit board or a data network, extending from the top to the bottom.

PROJET 10

OPENCLASSROOMS: FORMATION
INGÉNIEUR DATA



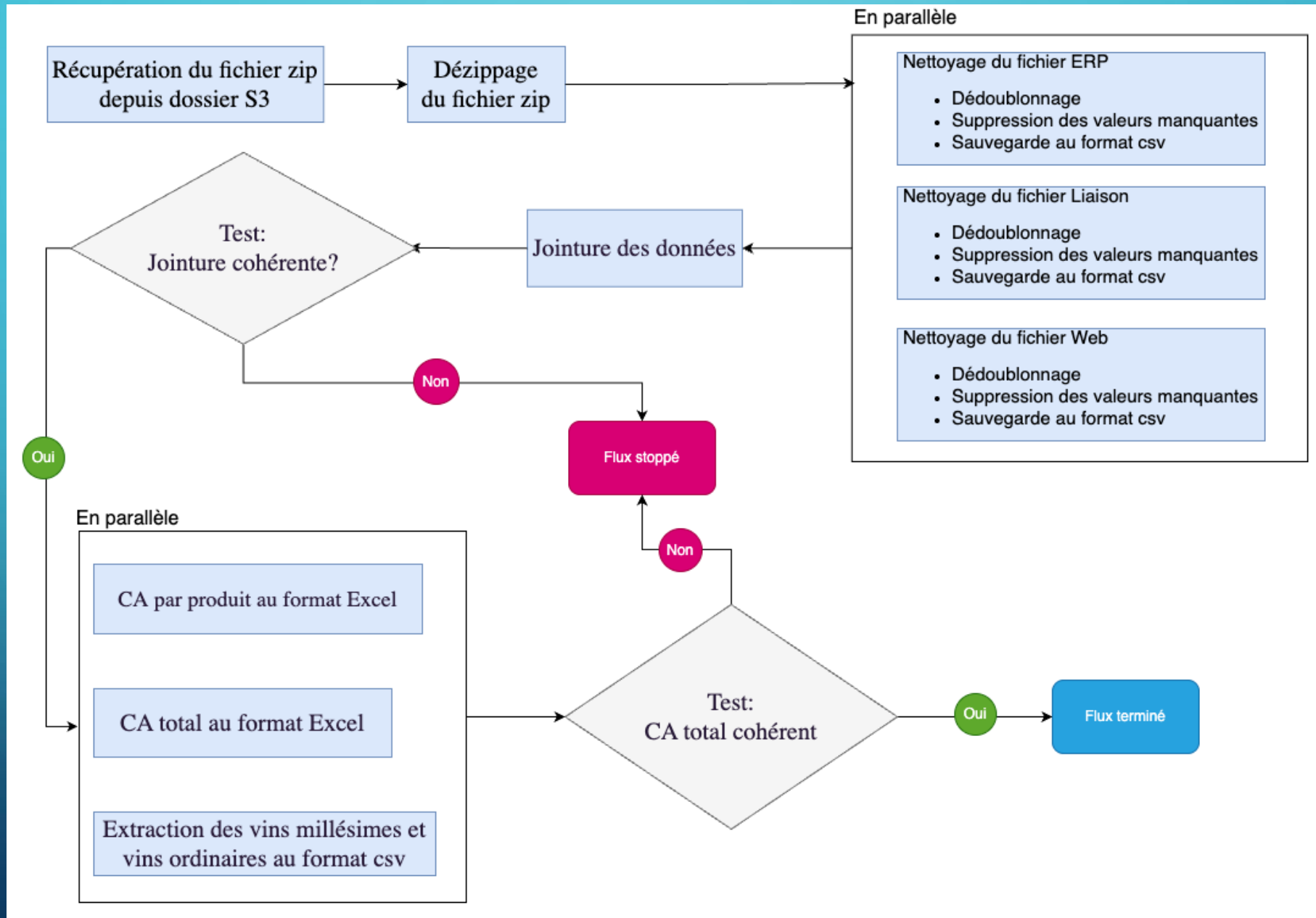
Contexte:

Data Engineer dans l'entreprise BottleNeck, un marchand de vin prestigieux.

But du projet:

Automatiser une chaîne de traitement et d'analyse de données.


DIAGRAMME DE FLUX





INSTALLATION DE KESTRA



```
[(base) nicolas@Mac-Nicolas Kestra % docker logs kestra
19:28:11.650 INFO main org.flywaydb.core.FlywayExecutor Database: jdbc:postgresql://postgres:5432/kestra (PostgreSQL 17.5)
19:28:11.144 INFO main o.f.core.internal.command.DbValidate Successfully validated 36 migrations (execution time 00:00.023s)
19:28:11.499 INFO main o.f.core.internal.command.DbMigrate Current version of schema "public": 1.38
19:28:11.486 INFO main o.f.core.internal.command.DbMigrate Schema "public" is up to date. No migration necessary.
19:28:12.507 INFO main i.kestra.core.plugins.PluginScanner Registered 99 core plugins (scan done in 81ms)
19:28:12.439 INFO standalone io.kestra.cli.AbstractCommand Starting Kestra 0.23.3 with environments [cli] [revision 1fb7943 / 2025-06-24T15:33]
19:28:13.175 INFO standalone i.kestra.core.plugins.PluginScanner Registered 637 plugins from 109 groups (scan done in 362ms)
19:28:13.533 INFO standalone io.kestra.cli.AbstractCommand Server Running: http://62c4241d014a:8080, Management server on port http://62c4241d014a:8081/health
```


INSTALLATION DE KESTRA






 Dashboards


 Flows


 Apps 



 Executions


 Logs


 Tests 


 Namespaces

 KV Store


 Secrets 

 Blueprints >

 Plugins

 Administration >


Flows


 Jump to...

 Ctrl/Cmd + K

 Import

 Source search

 Create

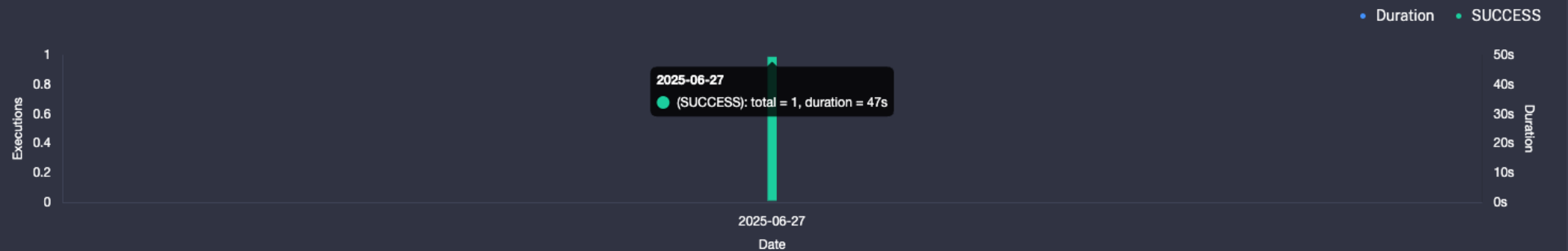
 scope= **USER** timeRange= **PT168H**







 Properties


Total Executions

Executions duration and count per date



Total: 1

<input type="checkbox"/> Id 	Labels	Namespace 	Last execution date	Last execution status	Triggers	Actions
<input type="checkbox"/> vins		bottleneck	Fri, Jun 27, 2025 6:31 PM	SUCCESS		

25 per page 

Total: 1

ÉDITION DU FLOW

127.0.0.1:8080/ui/main/flows/edit/bottleneck/vins/edit

Flows / bottleneck
vins ☆

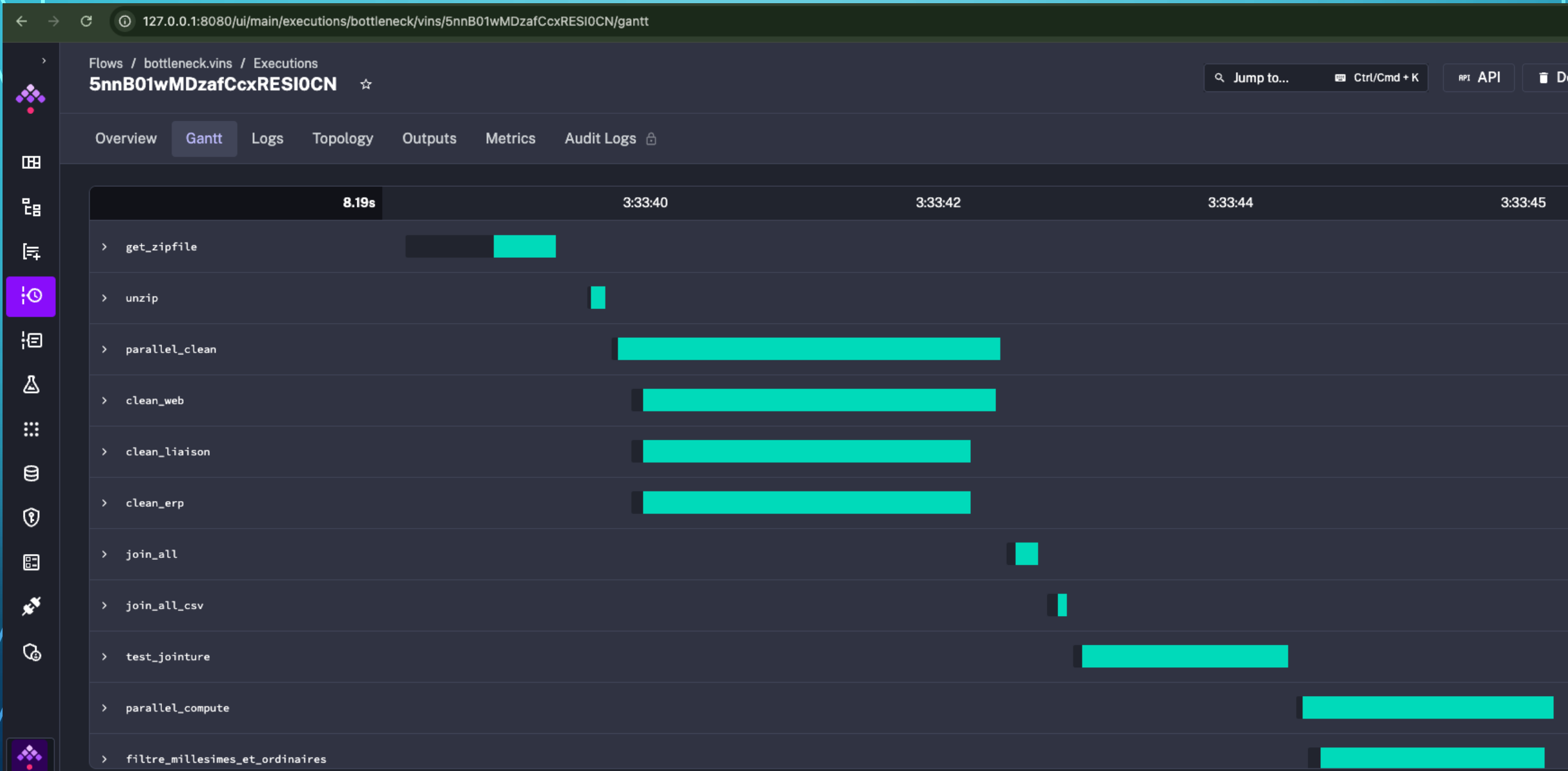
Overview Topology Executions **Edit** Revisions Triggers Logs Metrics Dependencies ⁰ Concurrency Audit Logs 🔒

<> Flow Code 📄 No-code 🏗️ Topology 📖 Documentation 📁 Files 📐 Blueprints

Flow Code

```
1 id: vins
2 namespace: bottleneck
3
4 tasks:
5   - id: get_zipfile
6     description: Récupération du fichier ZIP depuis un dossier S3. Ajout d'un retry (10 tentatives toutes les 6 minutes)
7     type: io.kestra.plugin.core.http.Download
8     uri: https://s3.eu-west-1.amazonaws.com/course.oc-static.com/projects/922_Data+Engineer/922_P10/bottleneck.zip
9     retry:
10       type: constant
11       maxAttempt: 10
12       interval: PT6M
13
14   - id: unzip
15     description: Dézippage du fichier zip
16     type: io.kestra.plugin.compress.ArchiveDecompress
17     algorithm: ZIP
18     from: "{{ outputs.get_zipfile.uri }}"
19
20   - id: parallel_clean
21     type: io.kestra.plugin.core.flow.Parallel
22     tasks:
23       - id: clean_erp
24         description: Suppression des doublons et des valeurs manquantes du fichier Erp
25         type: io.kestra.plugin.scripts.python.Script
26         containerImage: ghcr.io/kestra-io/pydata:latest
```

ÉXECUTION DU FLOW



NÉTTTOYAGE DES DONNÉES

- Utilisation de python et pandas
- Suppression des doublons et des valeurs manquantes sur chacun des fichiers Erp, Web et Liaison

TESTS SPÉCIFIQUES

- Utilisation de python et pandas
- Test de cohérence de la jointure: le nombre de lignes entre le fichier WEB et la jointure doit être identique
- Test de cohérence du chiffre d'affaires total: compris entre 10000 et 100000 euros

POUR ALLER PLUS LOIN

- Pour de plus gros projets, placer les scripts python sur **GITHUB**
=> allègement du fichier yaml
- Placer les fichiers csv et excel dans un bucket S3