MINISTÈRE
DE LA TRANSITION
ÉCOLOGIQUE
ET DE LA COHÉSION
DES TERRITOIRES

ECO LAB
Innovation - Écologie - Territoires

SERVICES DONNÉES

Présentation du code dans un projet Data

Amal YEFERNI et Nicolas ARIAS

15/03/2024

Plan

1. Projet indicateurs territoriaux de transition écologique

2. Problématique Data Science

3. Les métiers de la data

MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

ECO LAB
Innovation - Écologie - Territoires

1- Projet Indicateurs Territoriaux de la Transition Écologique

MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

Indicateurs Territoriaux de Transition Écologique

Les objectifs du projet

ECO LAB
Innovation - Écologie - Territoires

1) Partager les objectifs nationaux de la transition écologique au niveau des territoires.

1) Proposer aux collectivités territoriales un accès à des **indicateurs pertinents et reconnus de la transition écologique**, comparables et agrégeables entre territoires.

1) Création d'un **hub d'indicateurs** de la transition écologique.

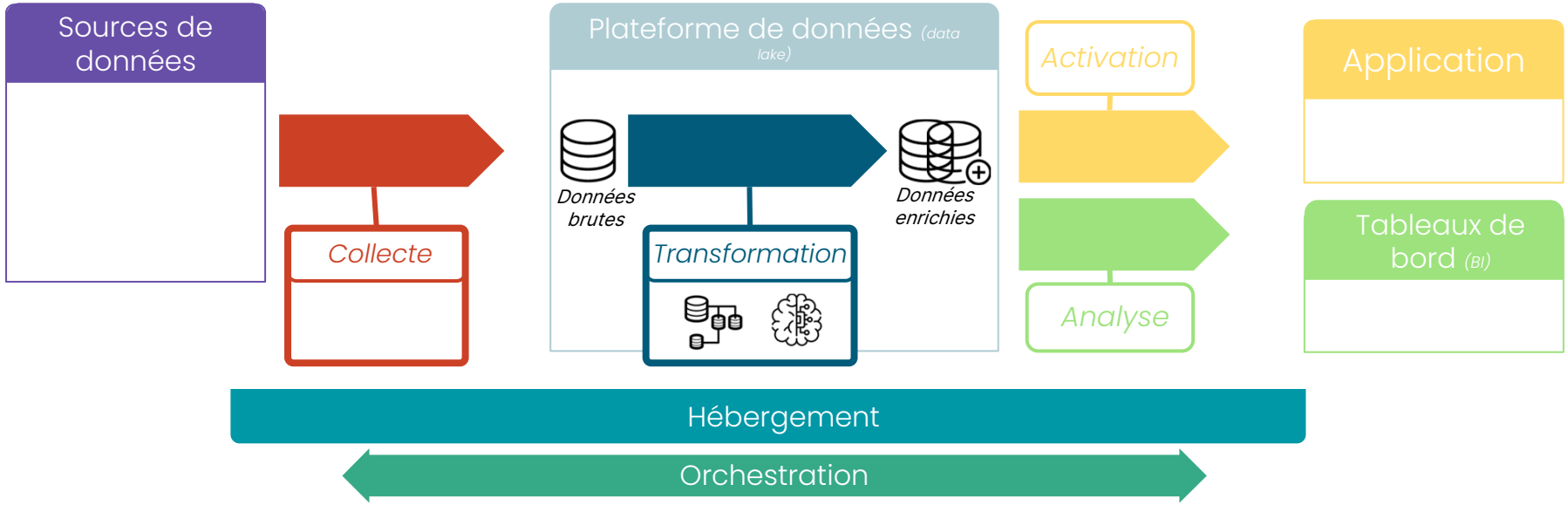
1) Les rendus du projet :
   - Création d'un outil ou intégration de ces indicateurs dans les bons outils déjà existant (Territoires en Transitions, Mon Espace Collectivité ..) pour **visualiser** et **exploiter** la donnée
   - Mise à disposition de ces indicateurs via API pour **récupérer** la donnée

MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

Indicateurs Territoriaux de Transition Écologique

Le flux de données

ECO LAB
Innovation - Écologie - Territoires



MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

Indicateurs Territoriaux de Transition Écologique

Type de données

ECO LAB
Innovation - Écologie - Territoires

Les données sont l'ensemble des informations générées par l'activité des différentes équipes.

Les deux principaux types de données :

Données structurées	Données non structurées
• S'affichent sous forme de rangée, colonne au sein d'une base de données • Exemple: dates,nombres (numéro de téléphone, numéro de sécurité sociale ..)	• Elles sont représentées sans un format prédéfini qui faciliterait leur accès et leur traitement. • Exemple : images, audios, vidéos, emails, fichiers pdf..

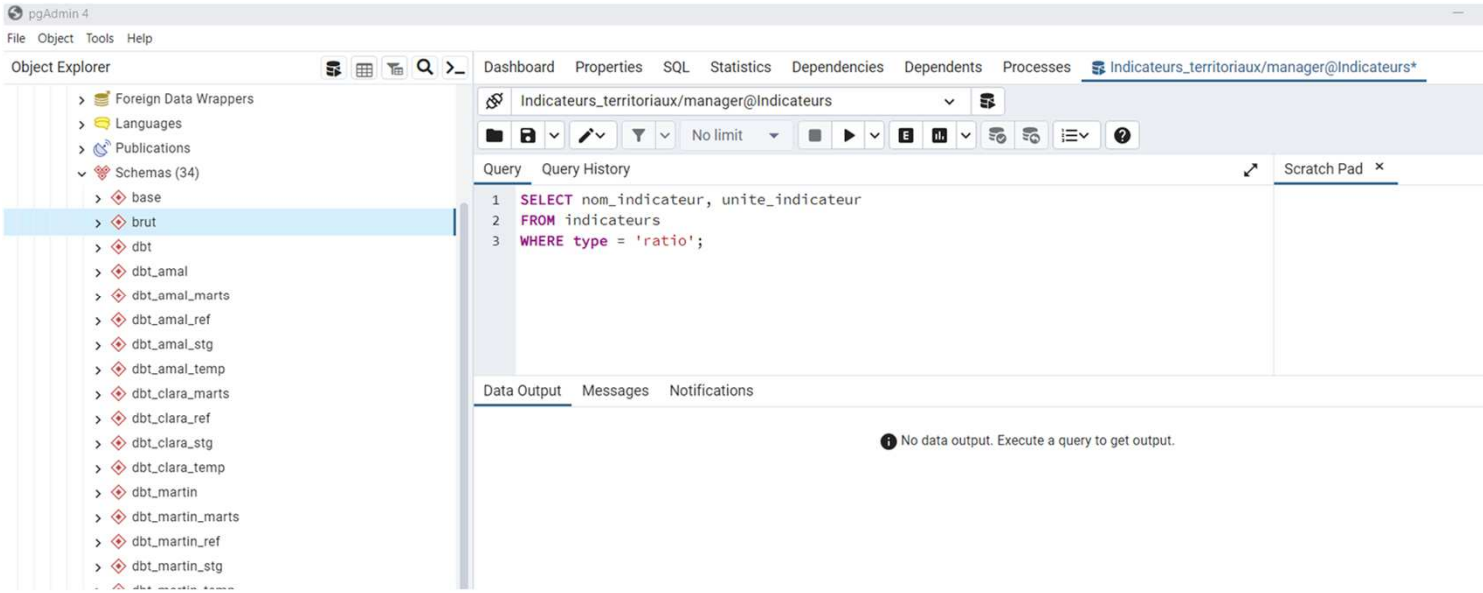
⇩

Cas du projet Indicateurs Territoriaux : **Données structurées** (excel, csv, json etc.)

Indicateurs Territoriaux de Transition Écologique

Récupération des données

- Données disponibles sous forme de fichiers (Excels, CSV), directement dans les bases de données, via des API
- Langage SQL



Indicateurs Territoriaux de Transition Écologique

Récupération des données

- Python permet aussi de lancer des requêtes SQL dans un script

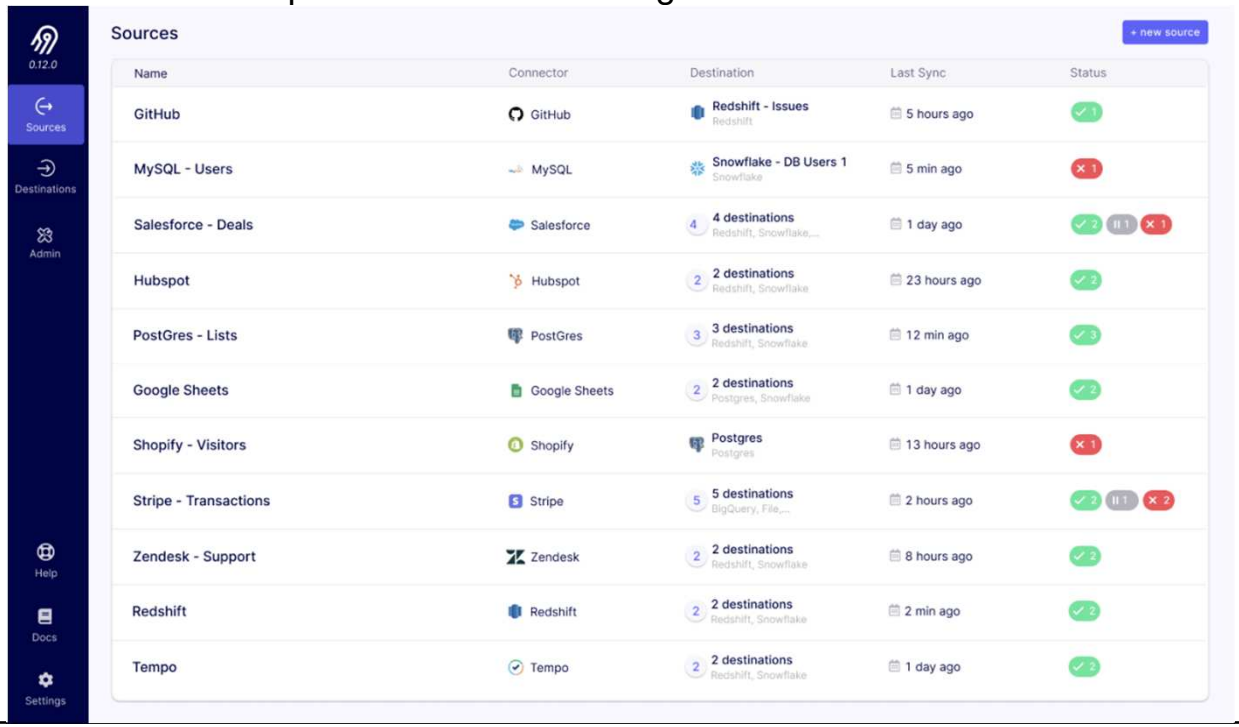
```python
import sqlalchemy

# instancier la connection à la base de données externe
engine = sqlalchemy.create_engine('postgresql://admin:password@indicateurs_source_externe/database')
connection = engine.connect()

# définir la table qui nous intéresse
indicateurs = sqlalchemy.Table("indicateurs")

# lancer une requête qui récupère les indicateurs de type ratio
query = sqlalchemy.select([indicateurs]).where(type = 'ratio')
result = connection.execute(query)
resultset = result.fetchall()

```

Récupération des données d'une source externe

Indicateurs Territoriaux de Transition Écologique

Récupération des données

- Automatisation de la récupération des données grâce à des outils no-code comme Airbyte



Indicateurs Territoriaux de Transition Écologique

Chargement des données dans la base de données

- Chargement des données dans la base de données à l'aide d'un script Python

```
17
18  import pandas as pd
19
20  # charger le fichier csv à importer
21  df = pd.read_csv("C:\Documents\indicateurs_a_importer.csv")
22
23  # instancier la connection à la base de données du projet indicateurs_territoriaux
24  engine = sqlalchemy.create_engine('postgresql://admin:password@indicateurs_territoriaux/database')
25
26  # charger les données dans la base de données
27  df.to_sql('indicateurs', engine)
28
```

Indicateurs Territoriaux de Transition Écologique

Chargement des données dans la base de données

- Données sous forme de tableaux dans la base de données

	codgeo_libelle text	libelle_variable text	variable text	libelle_sous_champ text	sous_champ text	type_var text	unite text	no_ind text
1	CC Faucigny-GliÃ¨res	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	30 Ã 39 ans	30_39	i	%	i001
2	CC Faucigny-GliÃ¨res	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	40 Ã 49 ans	40_49	i	%	i001
3	CC Faucigny-GliÃ¨res	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	50 Ã 59 ans	50_59	i	%	i001
4	CC Faucigny-GliÃ¨res	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	60 Ã 74 ans	60_74	i	%	i001
5	CC Faucigny-GliÃ¨res	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	75 ans et plus	75plus	i	%	i001
6	CC Faucigny-GliÃ¨res	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	moins de 30 ans	moins30	i	%	i001
7	CC Faucigny-GliÃ¨res	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	total	total	i	%	i001
8	CC du Pays de PontchÃ¢teau Saint-Gild...	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	30 Ã 39 ans	30_39	i	%	i001
9	CC du Pays de PontchÃ¢teau Saint-Gild...	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	40 Ã 49 ans	40_49	i	%	i001
10	CC du Pays de PontchÃ¢teau Saint-Gild...	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	50 Ã 59 ans	50_59	i	%	i001
11	CC du Pays de PontchÃ¢teau Saint-Gild...	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	60 Ã 74 ans	60_74	i	%	i001
12	CC du Pays de PontchÃ¢teau Saint-Gild...	Taux de pauvretÃ© monÃ©taire total et...	taux_pvt	75 ans et plus	75plus	i	%	i001
13	CC du Pays de PontchÃ¢teau Saint-Gild	Taux de pauvretÃ© monÃ©taire total et	taux_pvt	moins de 30 ans	moins30	i	%	i001

Indicateurs Territoriaux de Transition Écologique

Nettoyage des données

- Étape **indispensable** d'un projet data
- Sans nettoyage, les résultats des analyses risquent d'être faussés
- Cette étape permet d'apprendre à **connaître les données brutes**, en repérer **les points d'intérêt** et de **détecter les incohérences** au niveau des données
- Les problèmes dans les ensembles de données peuvent provenir (par exemple des fautes de saisies)

Avantages :

- ❖ Renforce la pertinence des données en réduisant les incohérences
- ❖ Améliore la fiabilité et la valeur des données

⇒ Permet une prise de décision mieux avisée et plus précise

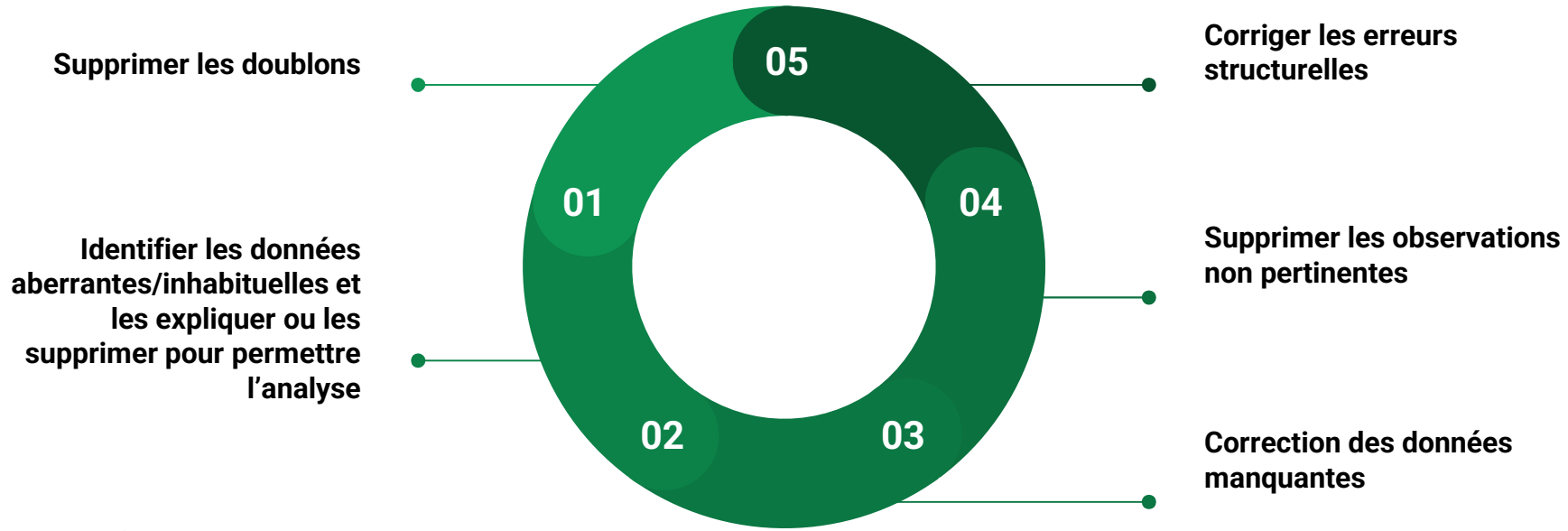
MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

Indicateurs Territoriaux de Transition Écologique

Nettoyage des données

ECO LAB
Innovation - Écologie - Territoires

Les principaux traitement des données :



L'étape de nettoyage se fait via des **scripts python/R** ou autres langages (voir script python)

Indicateurs Territoriaux de Transition Écologique

Transformation des données

Anonymisation des données pour masquer :

- Les données sensibles
- Les données permettant d'identifier une personne :
  - Noms
  - Addresses
  - numéros de téléphone etc.

⇒ Utilisation d'algorithme de **chiffrement**

```python
import pandas as pd
import hashlib

# Charger les données
data = {
    'Nom': ['Alice', 'Bob', 'Charlie', 'David'],
    'Age': [25, 30, 35, 40]
}

df = pd.DataFrame(data)

# Appliquer la fonction d'anonymisation à la colonne Nom
df['Nom_anonyme'] = df['Nom'].apply(lambda x : hashlib.sha256(x.encode()).hexdigest())

# Supprimer la colonne originale des noms
df.drop(columns=['Nom'], inplace=True)

print(df)
```

```
   Age                                        Nom_anonyme
0   25  3bc51062973c458d5a6f2d8d64a023246354ad7e064b1e...
1   30  cd9fb1e148ccd8442e5aa74904cc73bf6fb54d1d54d333...
2   35  6e81b1255ad51bb201a2b8afa9b66653297ae0217f833b...
3   40  a6b54c20a7b96eeac1a911e6da3124a560fe6dc042ebf2...
```

Indicateurs Territoriaux de Transition Écologique
Transformation des données

Ajout de nouvelles variables :

```
from datetime import datetime

df['date'] = pd.to_datetime(df['date'], format = "%d/%m/%Y")
df['annee'] = df['date'].dt.year
df.head()
```

	typezone	zone	codezone	date	nb_stations	nb_pdc	annee
0	Communes	L'Abergement-Clémenciat	1001	2021-02-05	1	0.0	2021
2	Communes	Ambérieu-en-Bugey	1004	2021-02-05	0	0.0	2021
3	Communes	Ambérieux-en-Dombes	1005	2021-02-05	1	2.0	2021
4	Communes	Ambléon	1006	2021-02-05	0	0.0	2021

Indicateurs Territoriaux de Transition Écologique

Transformation des données

Agrégation des données :

```
## Chaque commune est associée à un code EPCI qui lui correspond
```

	typezone	zone	codezone	date	nb_stations	nb_pdc	epci_code
84285	Communes	Saint-Laurent-en-Beaumont	38413	23/08/2021	0	0.0	200040657
84286	Communes	Sainte-Luce	38414	23/08/2021	0	0.0	200040657
84287	Communes	Saint-Marcel-Bel-Accueil	38415	23/08/2021	0	0.0	200068542
84288	Communes	Saint-Marcellin	38416	23/08/2021	2	4.0	200070431
84289	Communes	Sainte-Marie-d'Alloix	38417	23/08/2021	0	0.0	200018166

```
## calcul de nombre de stations de recharge de véhicules électriques pour les EPCI à partir des communes
df_total.groupby('epci_code')['nb_stations'].sum()
```

```
epci_code
200000925     0
200006682     0
200006971    27
200007052     6
200015162     4
             ..
247800550     0
248200016     0
248400251    24
248400335     6
ZZZ           0
Name: nb_stations, Length: 315, dtype: int64
```

Indicateurs Territoriaux de Transition Écologique

Analyse de données

Il est possible **d'extraire** de **l'information** à partir de la **donnée** de plusieurs manières et cela en fonction du **besoin**.

Exemple : Pour savoir combien il y a de 6 dans ce tableau :

6	9	5	3	9	6	1
2	6	4	4	2	4	7
5	4	6	4	1	6	8
6	8	4	1	4	4	9
4	7	6	5	6	6	5
3	4	2	7	5	4	2
5	8	7	8	5	3	6
4	9	9	6	8	6	10
3	10	6	3	2	4	3
9	4	3	6	4	8	6
6	5	4	9	1	3	0
3	7	0	9	9	9	3
5	7	2	0	6	10	6
7	9	1	1	2	3	4
6	3	9	7	4	1	6
9	10	7	4	1	1	5
3	10	3	5	3	9	8
7	7	6	8	6	2	4
9	9	0	6	7	0	6
9	1	7	2	8	6	9

Indicateurs Territoriaux de Transition Écologique
Analyse de données

Via une ligne de commande :

```
# Compter le nombre de 6 dans le tableau
nb_six = np.sum(matrice == 6)
```

```
nb_six
```

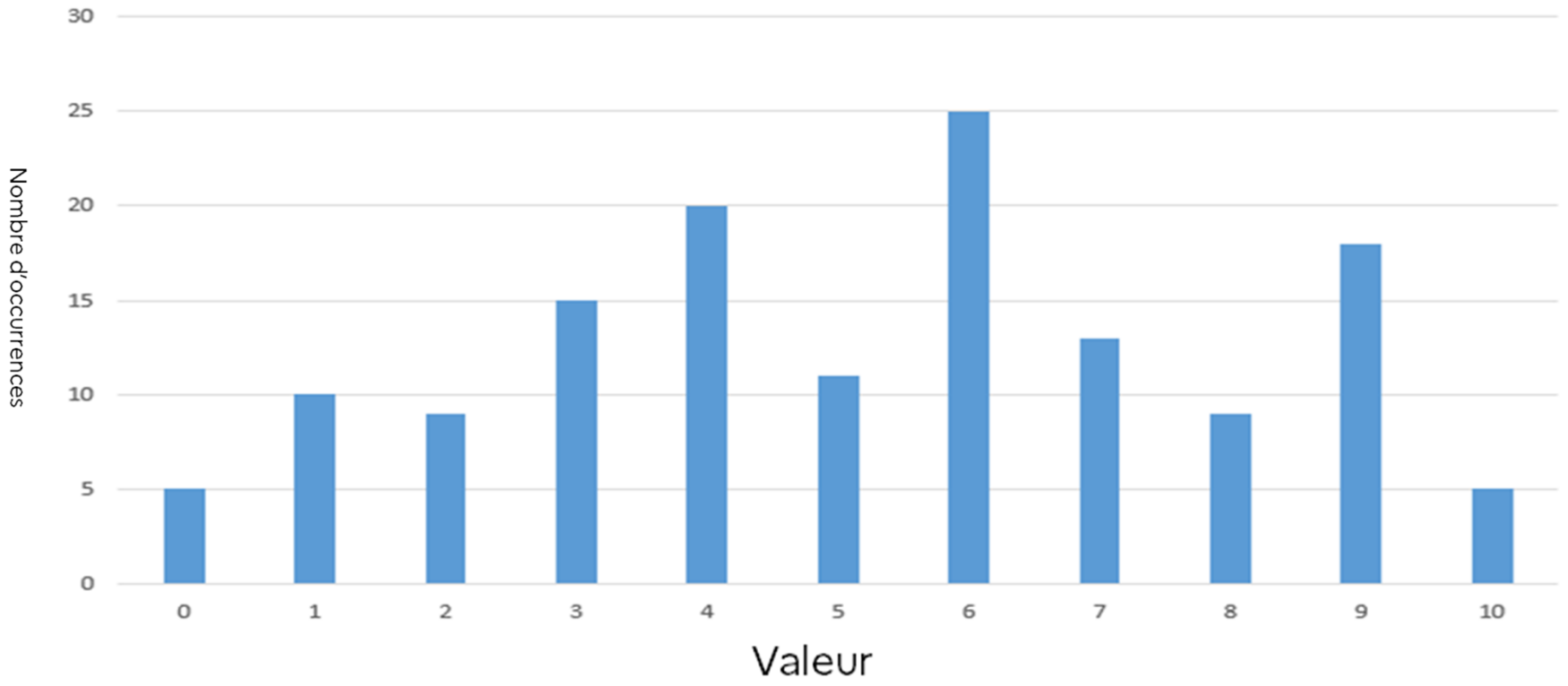
```
25
```

Via une représentation graphique qui se fait aussi avec une fonction prédéfinie de python :

```
import matplotlib.pyplot as plt

# Récupérer toutes les valeurs uniques de la matrice et leurs occurrences
valeurs_uniques, occurrences = np.unique(matrice, return_counts=True)

# Créer le graphe à barres
plt.bar(valeurs_uniques, occurrences)

# Afficher le graphe
plt.show()
```

Indicateurs Territoriaux de Transition Écologique

Analyse de données



MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

Indicateurs Territoriaux de Transition Écologique

Types de visualisation

ECO LAB
Innovation - Écologie - Territoires

COMPOSANTS GRAPHIQUES

Tableaux	Graphes	Cartes
Ils restituent une **vision synthétique, bien qu'assez brute**, des données.	Ils restituent une **interprétation spécifique** de données et peuvent notamment être mobilisés pour illustrer un rapport écrit. Exemples : diagramme en bâtons, diagramme circulaire.	Elles mettent en avant la **dimension géographique** des données. Elles permettent notamment l'interprétation des tendances et des comparaisons régionales.

TYPES DE DATA VIS'

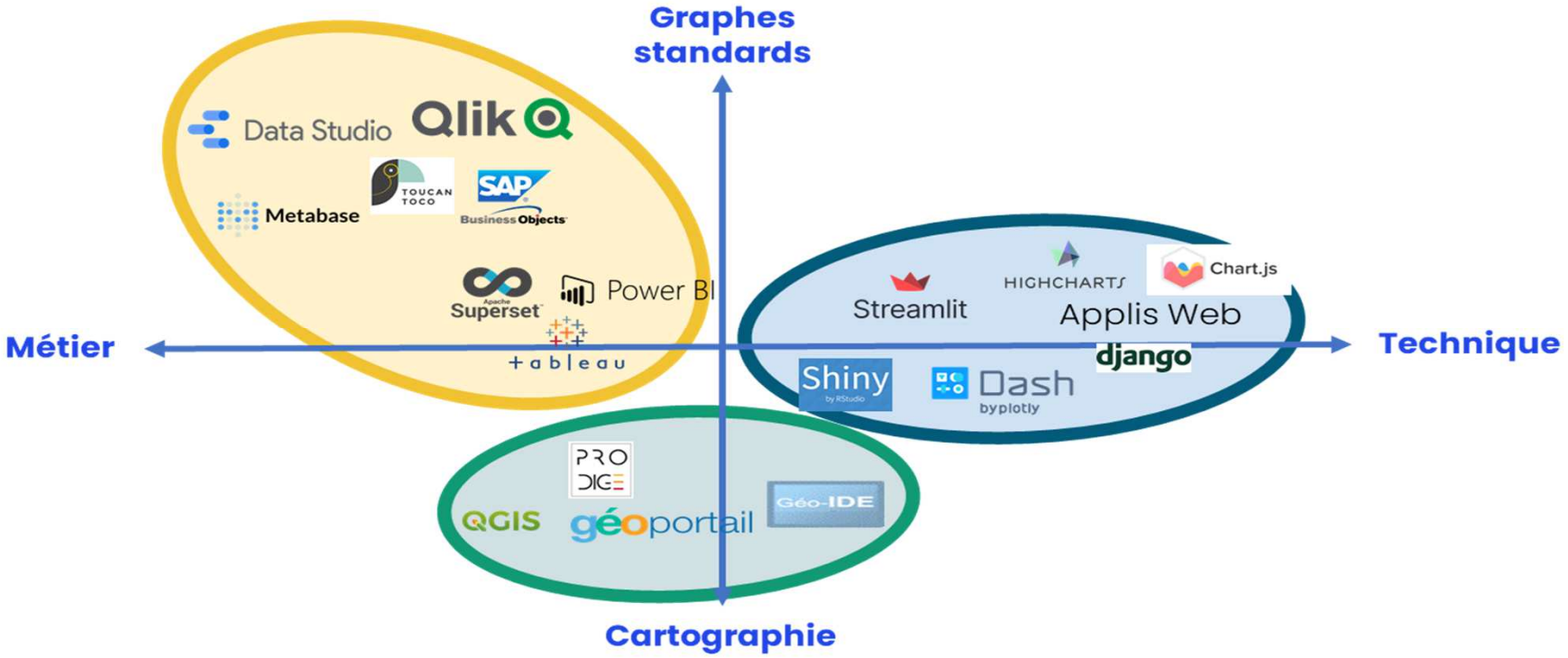
Rapports	Cartographies	Tableaux de bord	Applications interactives	Infographies

MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

Indicateurs Territoriaux de Transition Écologique

Outils de visualisations

ECO LAB
Innovation - Écologie - Territoires

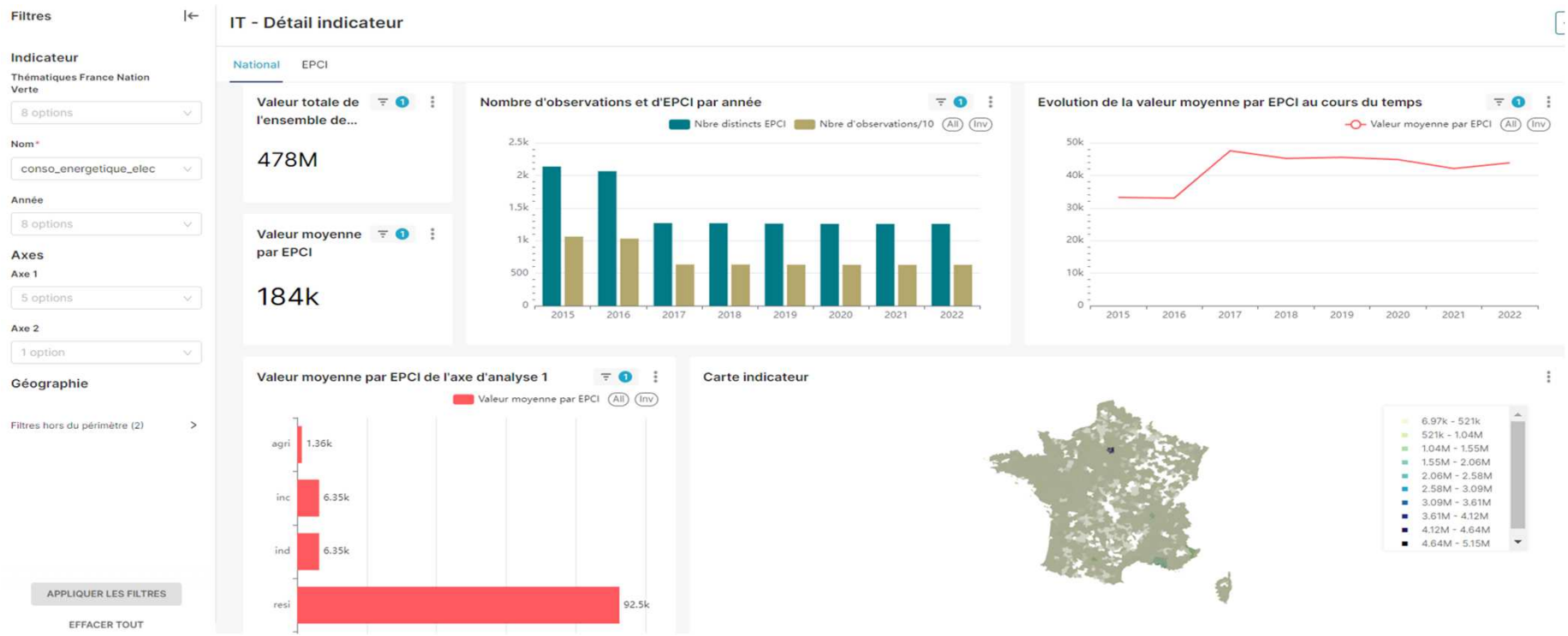


MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

Indicateurs Territoriaux de Transition Écologique

Tableau de bord - Superset

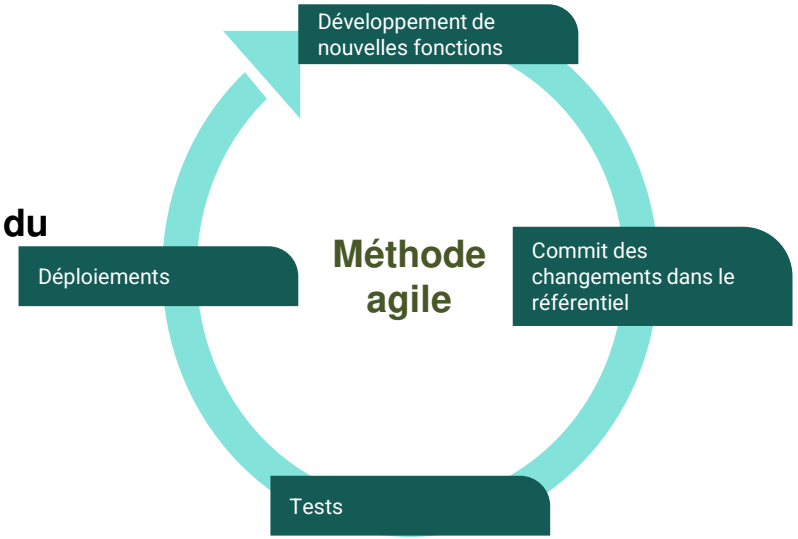
ECO LAB
Innovation - Écologie - Territoires



Indicateurs Territoriaux de Transition Écologique

Test du code

- **Détection précoce des erreurs** avant le déploiement
- Amélioration de la **qualité globale** du code en augmentant sa fiabilité
- **Confiance** dans le code
- Facilitation de la **maintenance** et de **l'amélioration du code** (intégration et déploiement en continue)



Indicateurs Territoriaux de Transition Écologique

Orchestration

- Besoin de lancer les étapes d'un pipeline dans le bon ordre et à la bonne **fréquence**
- Monitoring pour s'assurer que tout se passe bien

```
with DAG('example_dag', default_args=default_args, schedule_interval='@daily') as dag:
    start_task = DummyOperator(task_id='start')
    end_task = DummyOperator(task_id='end')
    start_task >> end_task
```

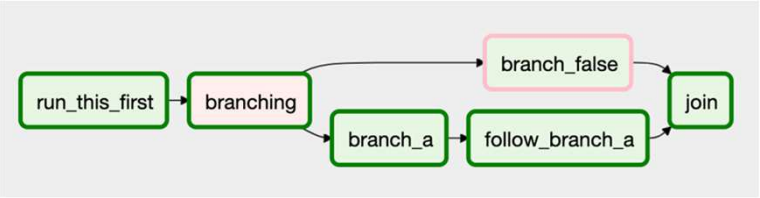


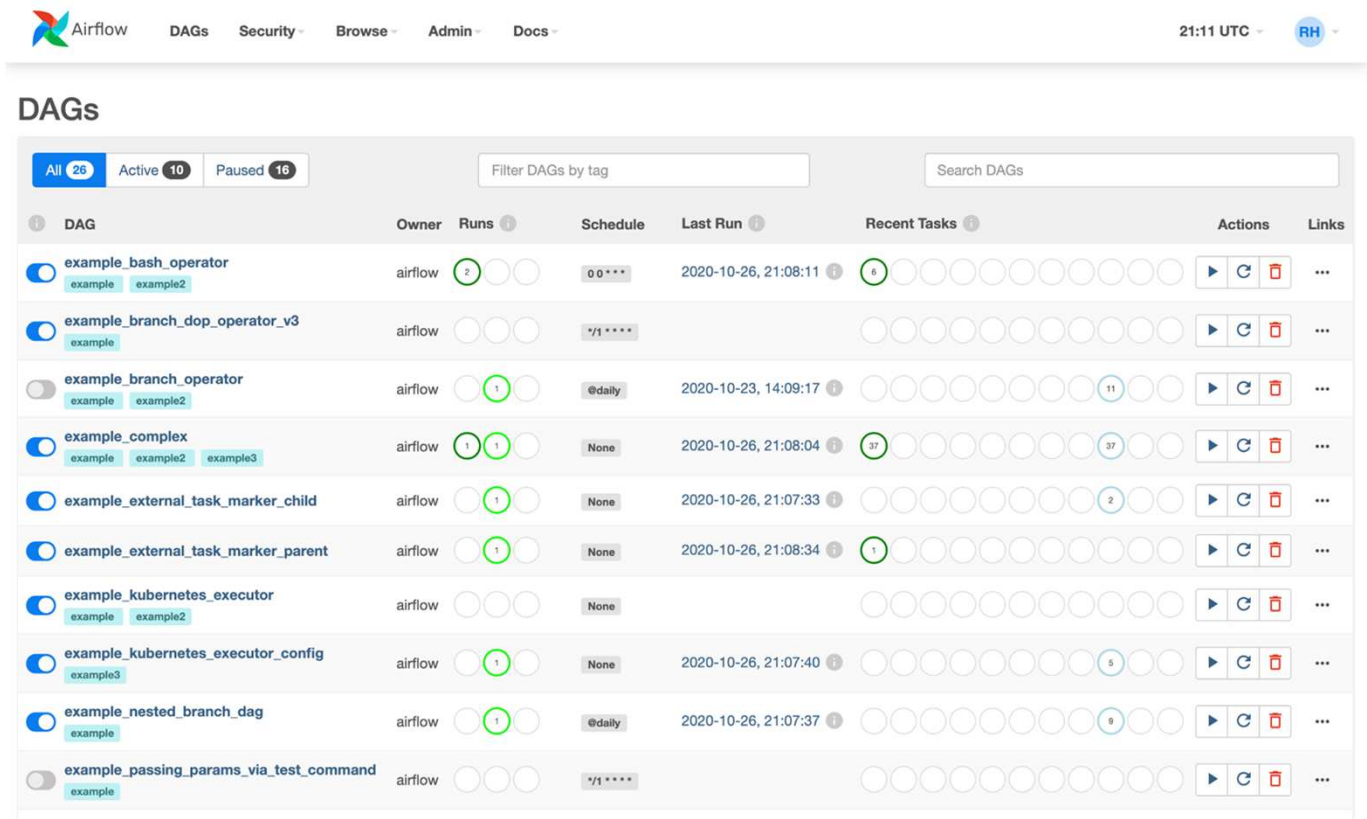
Diagramme Airflow

MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

Indicateurs Territoriaux de Transition Écologique

Orchestration

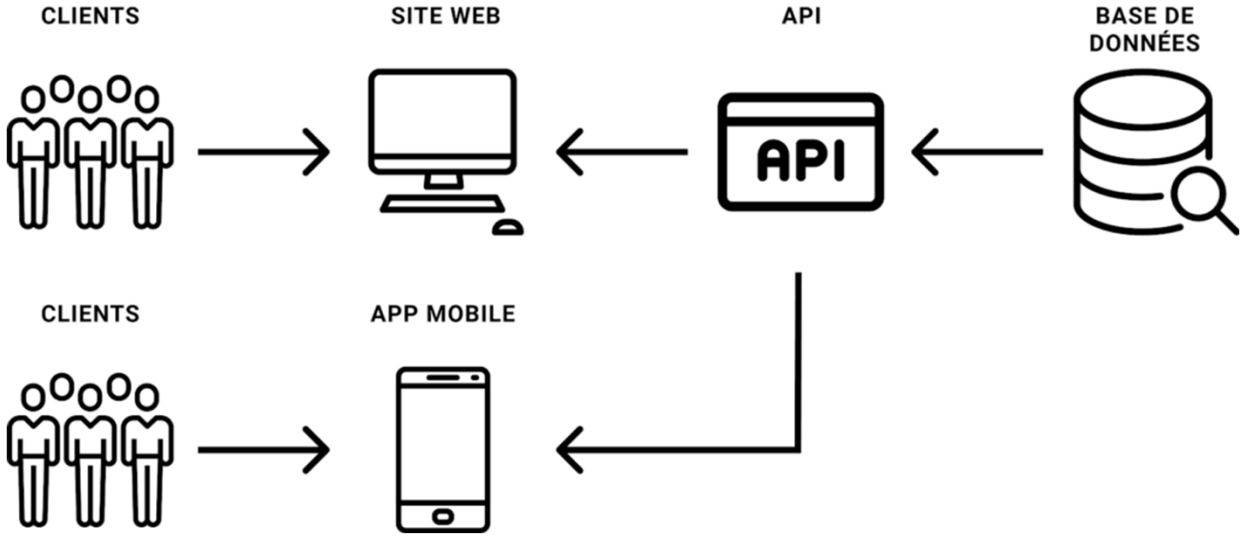
ECO LAB
Innovation - Écologie - Territoires



Indicateurs Territoriaux de Transition Écologique

Mise à disposition des données via API

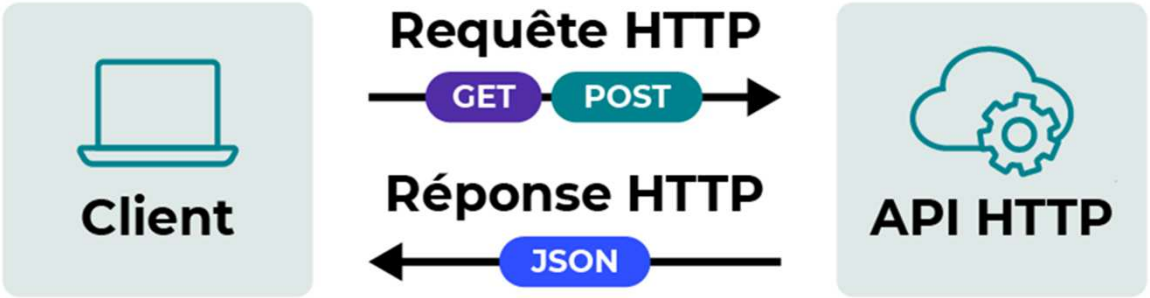
- Application Programming Interface (interface de programmation d'applications)



Source : CosaVostra

Indicateurs Territoriaux de Transition Écologique

Mise à disposition des données via API



Source : OpenClassroom

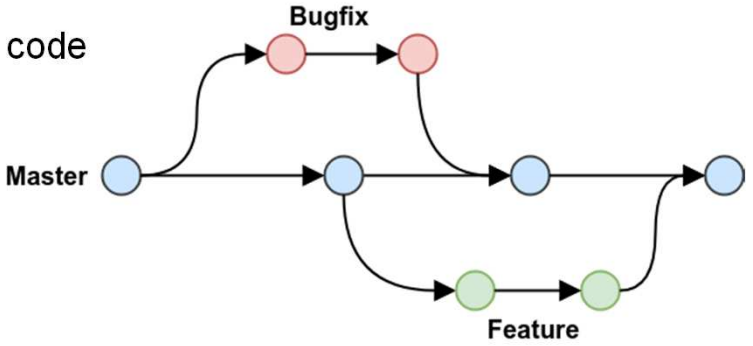
```
GET /indicateurs?departement=33 HTTP/1.1
Host: api.indicateurs_territoriaux.fr
```

```json
[
    {
        "indicateur" : "emissions_co2",
        "valeur" : "12",
        "unite" : "tonnes"
    },
    {
        "indicateur" : "longueur_pistes_cyclables",
        "valeur" : "210",
        "unite" : "km"
    }
]
```

Indicateurs Territoriaux de Transition Écologique

Versioning et GIT

- Collaborer entre plusieurs développeurs
  - Fusionner le code de 2 développeurs travaillant sur le même fichier
  - Développer plusieurs fonctionnalités en même temps
- Garder l'historique des modifications du code

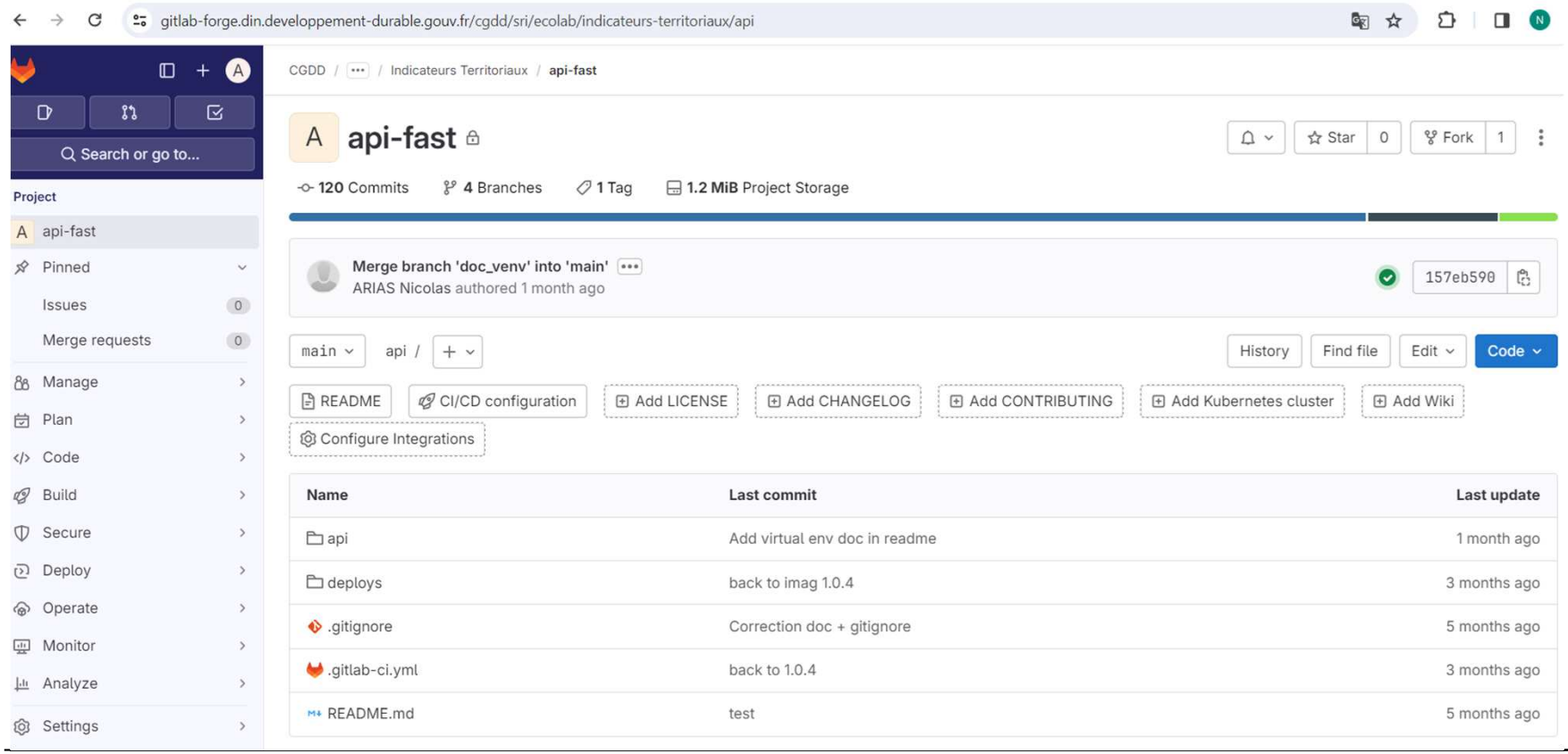


MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

Indicateurs Territoriaux de Transition Écologique

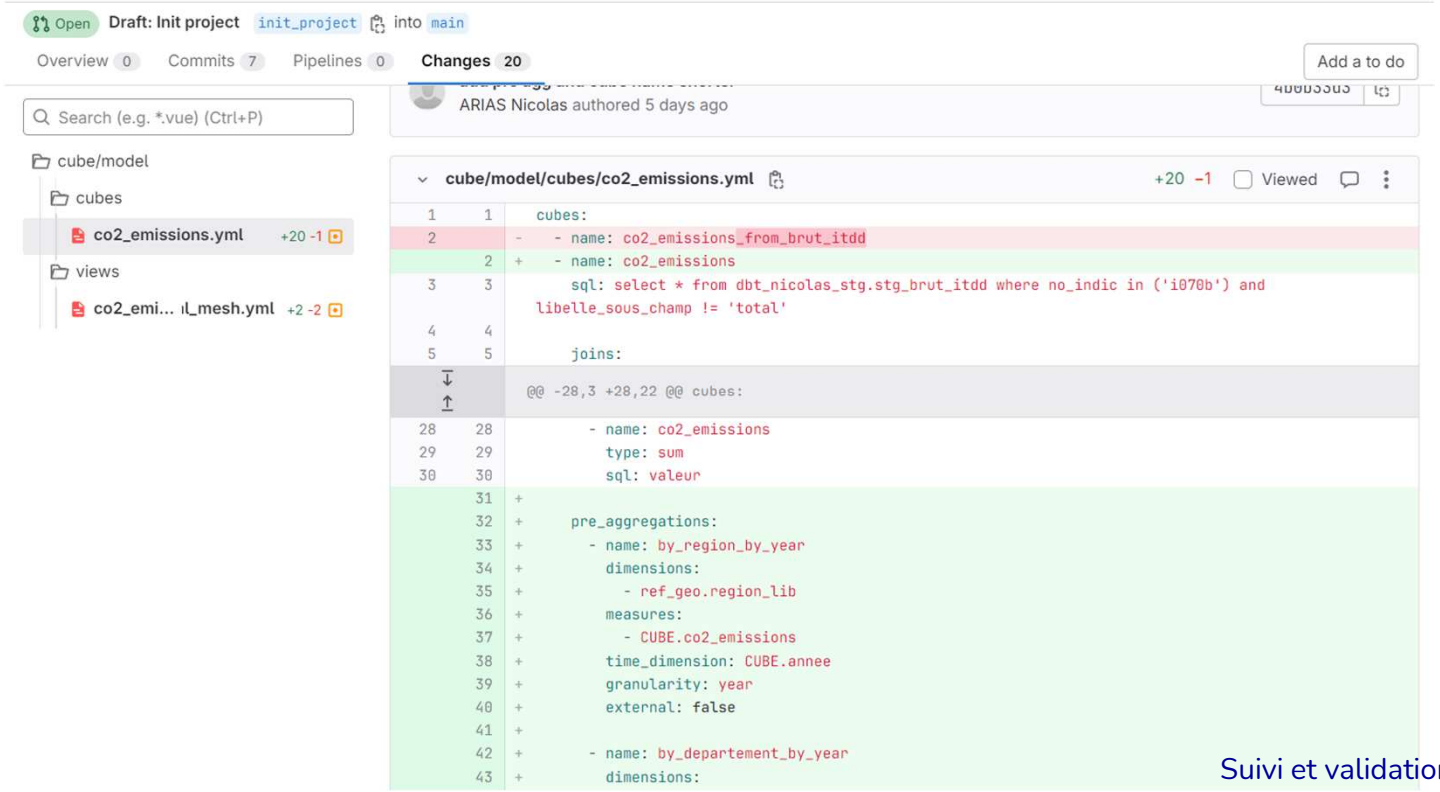
Versioning et GIT

ECO LAB
Innovation - Écologie - Territoires



Indicateurs Territoriaux de Transition Écologique

Versioning et GIT



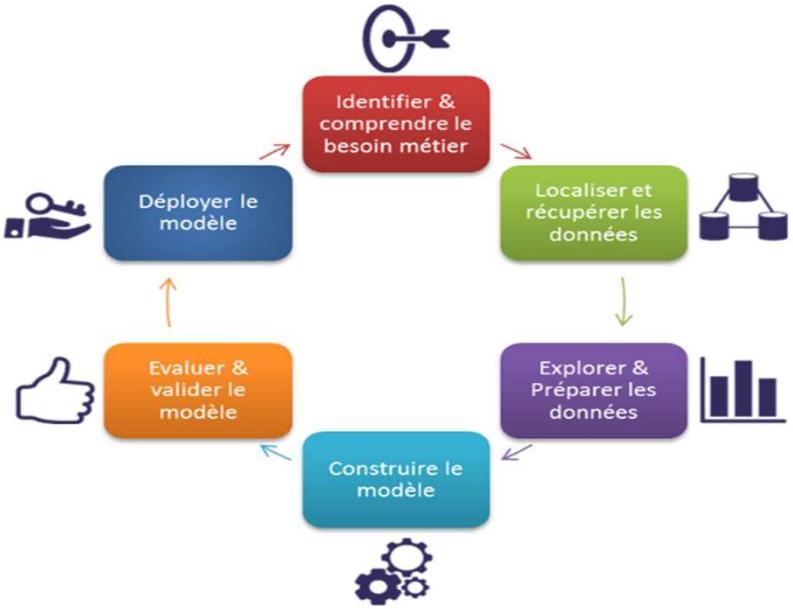
Suivi et validation des modifications

2. Problèmatique Data Science

Problématique Data Science

Etapes projet data science

Quelque soit la nature ou le domaine d'application d'un projet de données, les étapes du cycle de la vie de la données sont généralement similaires.



Problématique Data science

Exemple

	Ville	Longitude	Latitude
0	Paris	2.3522	48.8566
1	Marseille	5.3698	43.2965
2	Lyon	4.8357	45.7640
3	Toulouse	1.4442	43.6047
4	Nice	7.2657	43.7102
5	Nantes	-1.5531	47.2184
6	Strasbourg	7.7521	48.5734
7	Montpellier	3.8767	43.6109
8	Bordeaux	-0.5792	44.8378
9	Lille	3.0573	50.6292

On souhaite regrouper les villes en fonction de leur **proximité géographique en 3 groupes homogènes.**

Projet Data science

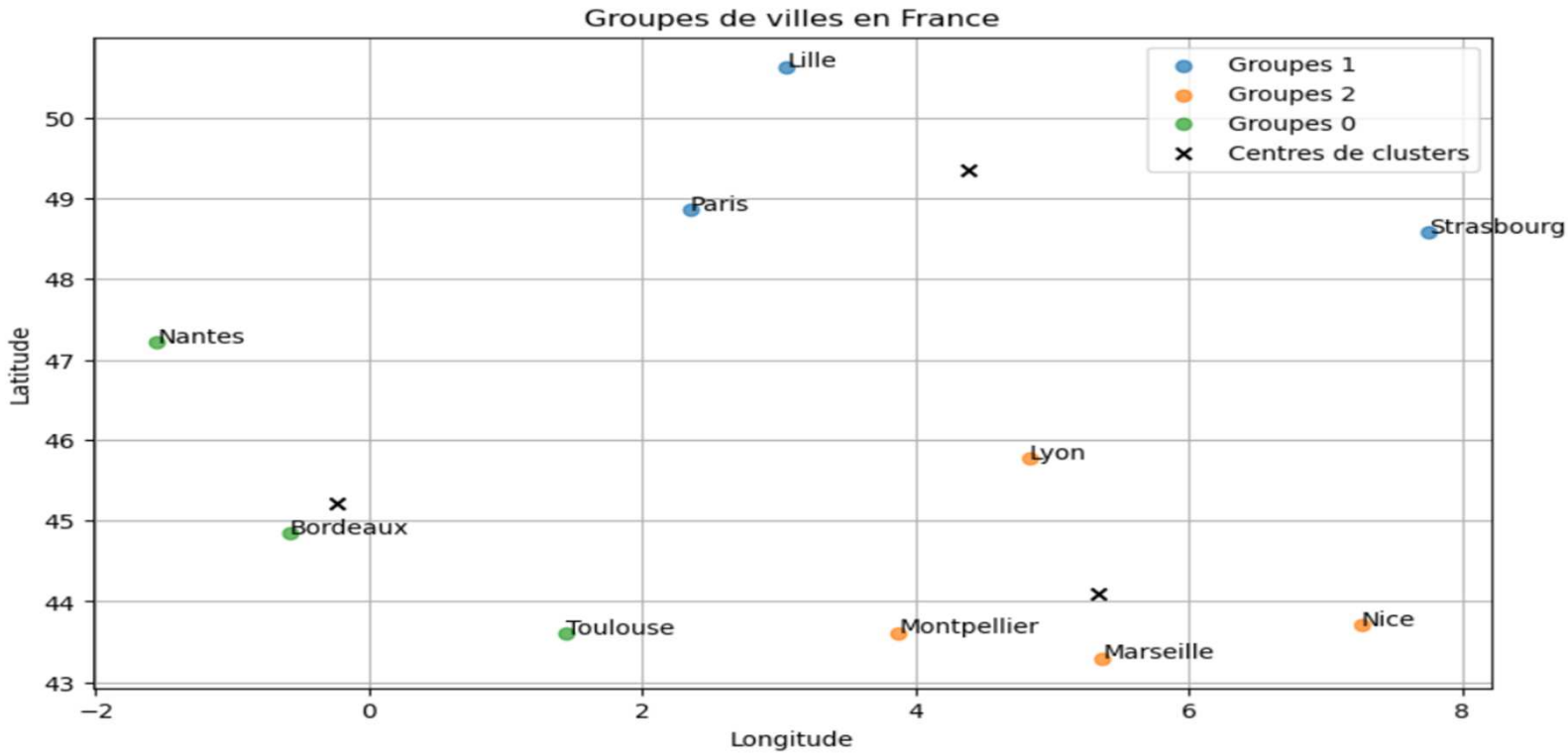
Implémentation de K-means


```python
# Appliquer l'algorithme K-means avec K=3
kmeans = KMeans(n_clusters=3, random_state=42)
df['Groupes'] = kmeans.fit_predict(df[['Longitude', 'Latitude']])

# Afficher les résultats
print(df)
```

```
         Ville  Longitude  Latitude  Groupes
0        Paris     2.3522   48.8566        1
1    Marseille     5.3698   43.2965        2
2         Lyon     4.8357   45.7640        2
3     Toulouse     1.4442   43.6047        0
4         Nice     7.2657   43.7102        2
5       Nantes    -1.5531   47.2184        0
6   Strasbourg     7.7521   48.5734        1
7  Montpellier     3.8767   43.6109        2
8     Bordeaux    -0.5792   44.8378        0
9        Lille     3.0573   50.6292        1
```

Projet Data science

Résultat K-means



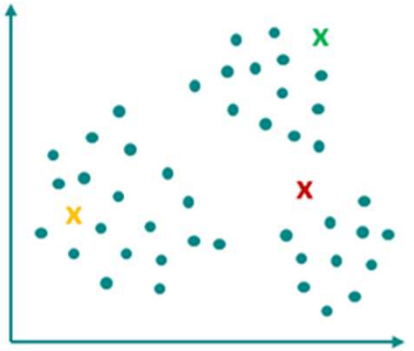
Projet Data science

Algorithme K-means

Nous avons utilisé un framework déjà existant . Cependant, nous pouvons aussi coder le K-means from scratch.

Le principe de K-means :

1. Choisir le nombre de groupes (dans notre cas 3).
2. L'algorithme sélectionne 3 points au hasard dans l'ensemble de données, appelés centroids, qui serviront de centres initiaux des groupes.



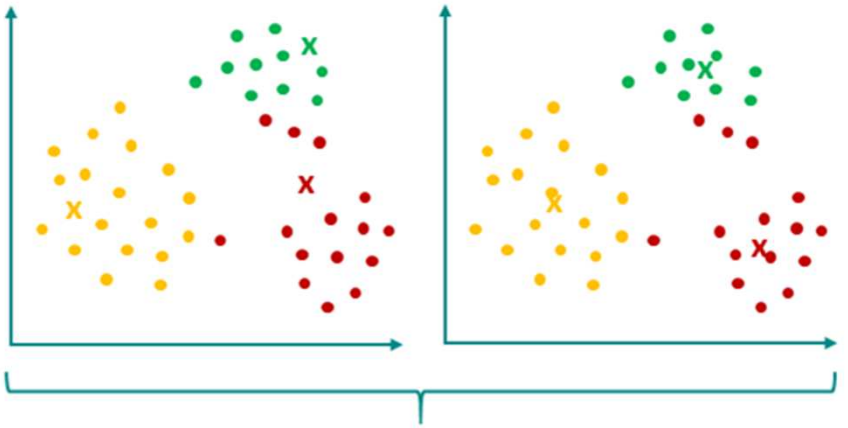
Projet Data science

Algorithme K-means

3. Il attribue chaque ville au groupe dont le centroide est le plus proche en termes de distance

4. Il déplace ensuite les centroids vers le centre de gravité de leurs groupes respectifs

3 4

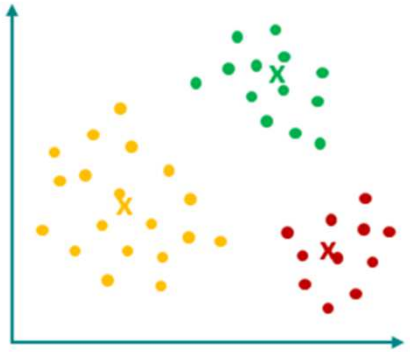


L'étape 3 et 4 sont répétées jusqu'à ce que les centroids ne se déplacent plus.

Projet Data science

Algorithme K-means

5. Chaque ville est donc maintenant associée à un groupe.



Une fois que le fonctionnement de l'algorithme a été assimilé, nous pouvons le coder from scratch.

Projet Data science

Algorithme K-means


```python
def k_means(X, K, max_iters=100):
    # Initialisation aléatoire des centroids
    centroids = X[np.random.choice(len(X), K, replace=False)]

    for _ in range(max_iters):
        # Assignation des points à leurs clusters les plus proches
        clusters = [[] for _ in range(K)]
        for point in X:
            closest_centroid_idx = np.argmin(np.linalg.norm(point - centroids, axis=1))
            clusters[closest_centroid_idx].append(point)

        # Mise à jour des centroids
        for i in range(K):
            if clusters[i]:
                centroids[i] = np.mean(clusters[i], axis=0)

    return centroids, clusters
```

MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

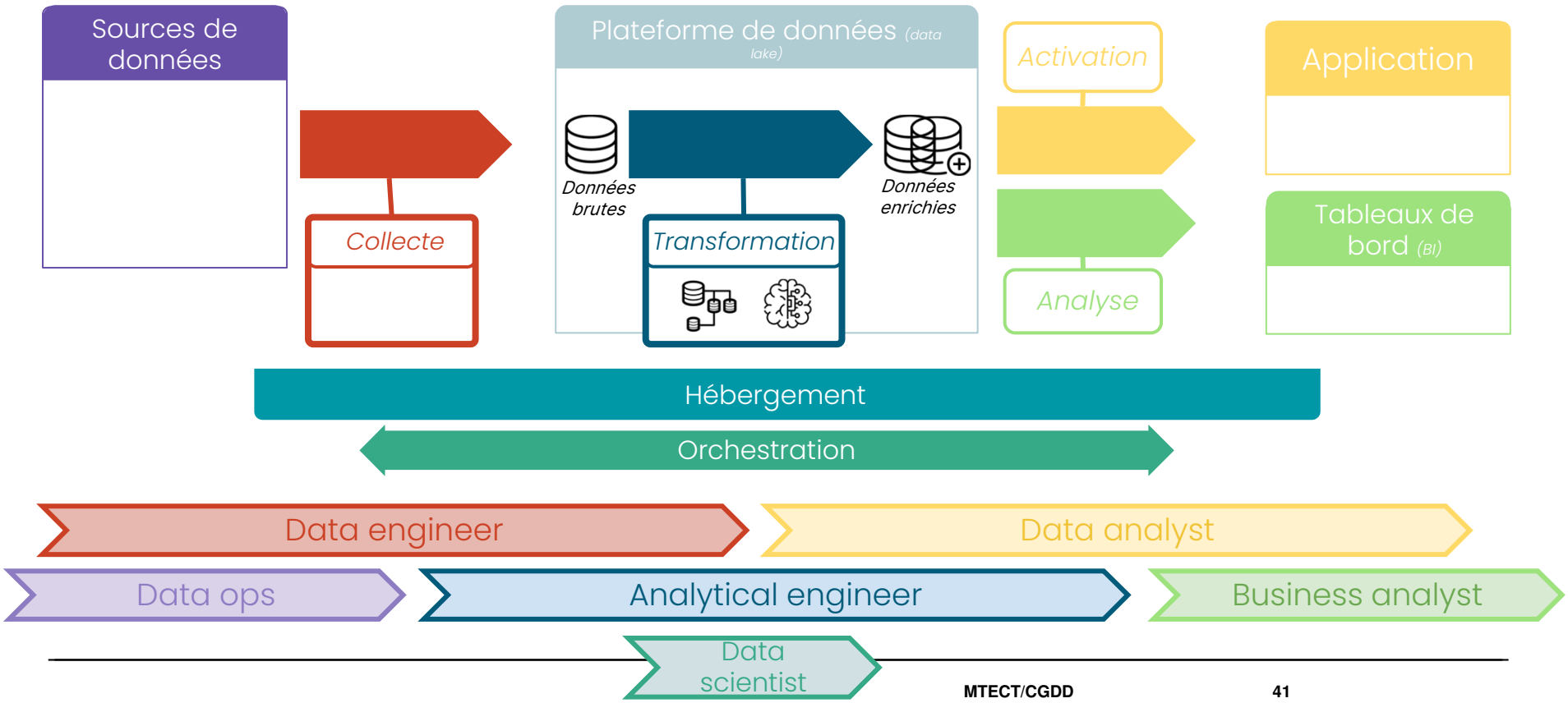
ECO LAB
Innovation - Écologie - Territoires

3. Les métiers de la data

MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

Les différents métiers de la Data

ECO LAB
Innovation - Écologie - Territoires



MINISTÈRES
TRANSITION ÉCOLOGIQUE
COHÉSION DES TERRITOIRES
TRANSITION ÉNERGÉTIQUE
MER
Liberté
Égalité
Fraternité

ECO LAB
Innovation - Écologie - Territoires

Merci !
Des questions ?