

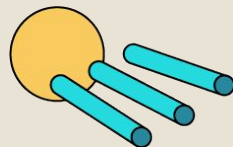
# Itinéraire de vacances

OCT24 - BOOTCAMP - Data Engineering

Olivier AZEAU  
Mathieu PAUL  
Samuel MOCHER



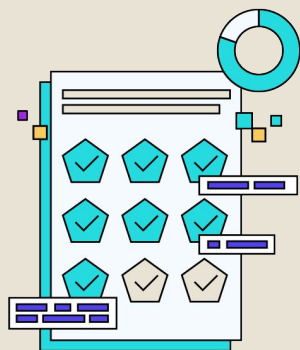
DataScientest





# SOMMAIR

E



- 1 Schéma Application / Récolte des données
- 2 Stockage / Normalisation / Graph
- 3 Consommation
- 4 Orchestration / Monitoring
- 5 Améliorations / évolutions possibles
- 6 Démonstration



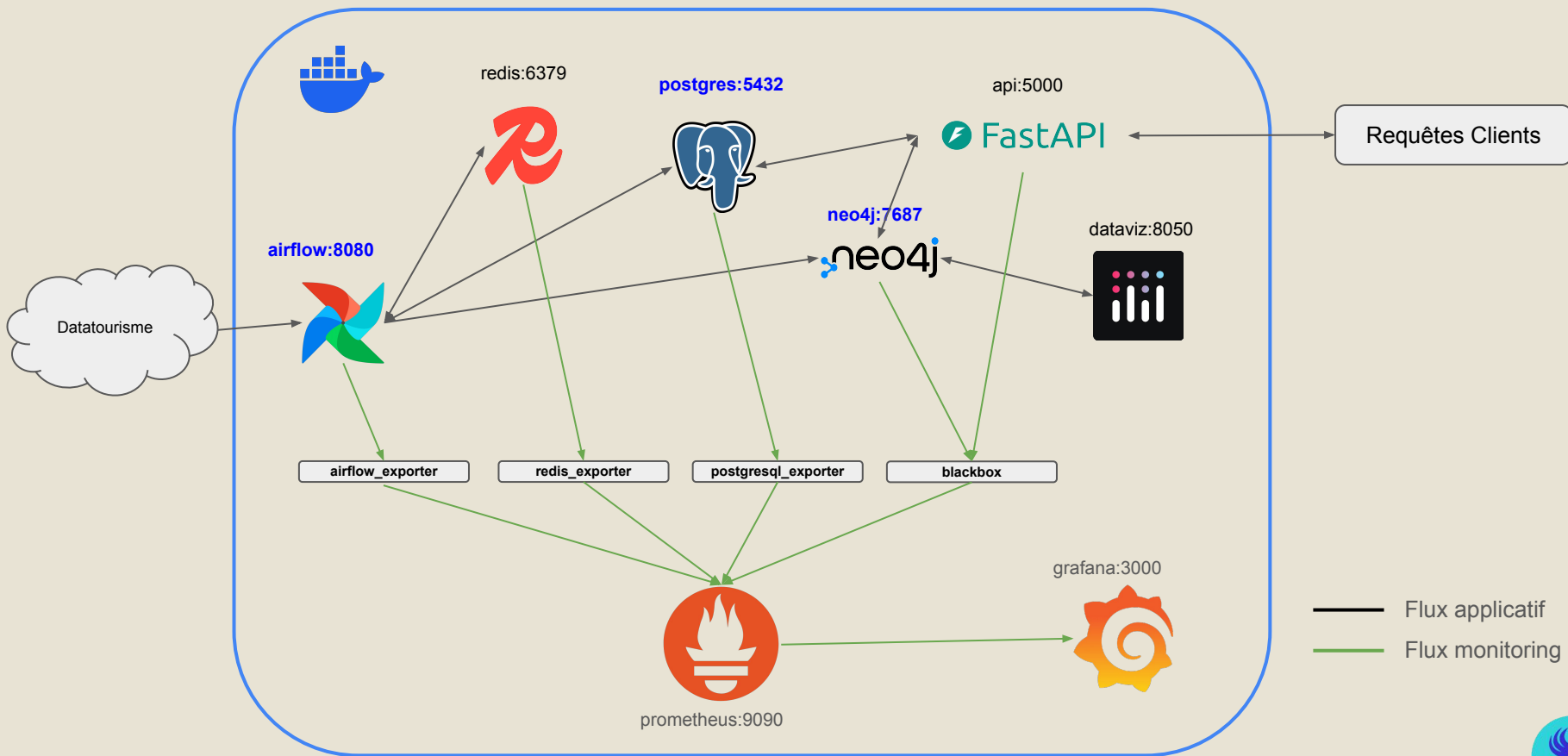
# Présentation du projet



- L'objectif du projet est la création d'une application permettant de proposer un itinéraire selon certains critères.
- L'utilisateur de l'application choisit des zones / points d'intérêt à visiter lors de son prochain voyage, lui propose un itinéraire détaillé optimisant son temps de voyage et de séjour.



## Schéma Applicatif




# Récolte des données

## Exploration



- Base de données publique
- Possibilités de ciblage du flux
  - Zone géographique
  - Types de POI
- Update toutes les 24h
- Disponible sous de multiples formats
  - JSON / XML / ...

Accueil Flux Applications Support

Accueil > Flux > Flux France Entiere + DOM TOM

### Flux : Flux France Entiere + DOM TOM

Paramètres Éditeur de requête Téléchargement Historique Administration

**Nom \***

Flux France Entiere + DOM TOM

**Description \***

Flux complet


**Format \***

Fichiers JSON

Les formats proposés dépendent du type de requête SPARQL (SELECT ou CONSTRUCT). [Voir un exemple du format sélectionné.](#)

**Webservice**

`https://diffuseur.datatourisme.fr/webservice/dc1a6e04c5f1c8825d1b081a593389d5/{app_key}`

 Pour consommer le flux à partir de l'une de vos applications, vous devez utiliser l'adresse ci-dessus en remplaçant le paramètre **{app\_key}** par la clé API dédiée à l'application, que vous trouverez dans la section **Applications**

Enregistrer



```

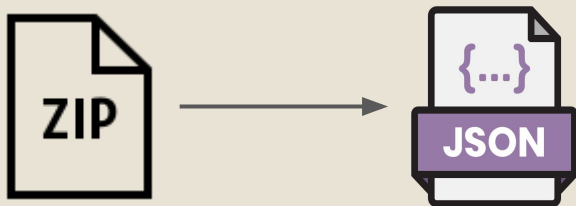
},
{
  "label": "Les Glénans - École de voile",
  "lastUpdateDatatourisme": "2024-04-11T01:15:27.212Z",
  "file": "0/00/38-0027aa71-bfb4-3f2f-b2d1-8e7c01164723.json"
},
{
  "label": "Chapelle de Lothéa",
  "lastUpdateDatatourisme": "2022-12-09T02:09:22.555Z",
  "file": "0/00/38-0028783f-43ea-3215-b8de-3def30c372bd.json"
},
{
  "label": "Chapelle de Kergrist",
  "lastUpdateDatatourisme": "2023-08-05T01:29:27.423Z",
  "file": "0/00/38-002b790b-dedd-363f-af5e-bb2e77b5aede.json"
},
{

```

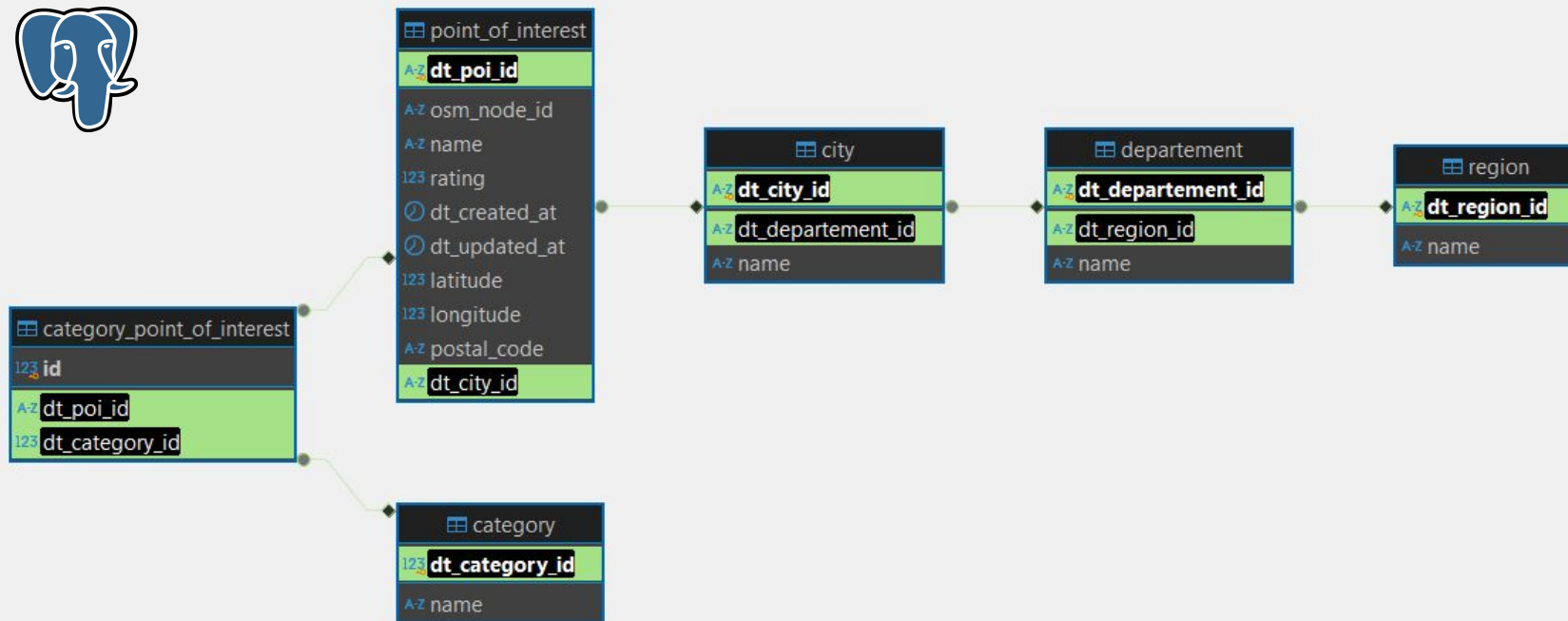
```

{
  "@id": "https://data.datatourisme.fr/38/0027aa71-bfb4-3f2f-b2d1-8e7c01164723",
  "dc:identifiant": "ASCBRE029FS00024",
  "@type": [
    "schema:Product",
    "PlaceOfInterest",
    "PointOfInterest",
    "Product",
    "SportsAndLeisurePlace"
  ],
  "rdfs:label": {
    "en": [
      "Les Glénans - École de voile"
    ],
    "fr": [
      "Les Glénans - École de voile"
    ]
  },
  "hasBeenCreatedBy": {
    "@id": "https://data.datatourisme.fr/3be8ff4f-e15a-3cb9-9a15-c9a135e38f89",
    "schema:legalName": "Finistère 360° - Direction",
    "@type": [
      "schema:Organization",
      "foaf:Agent",
      "Agent"
    ]
  }
}

```

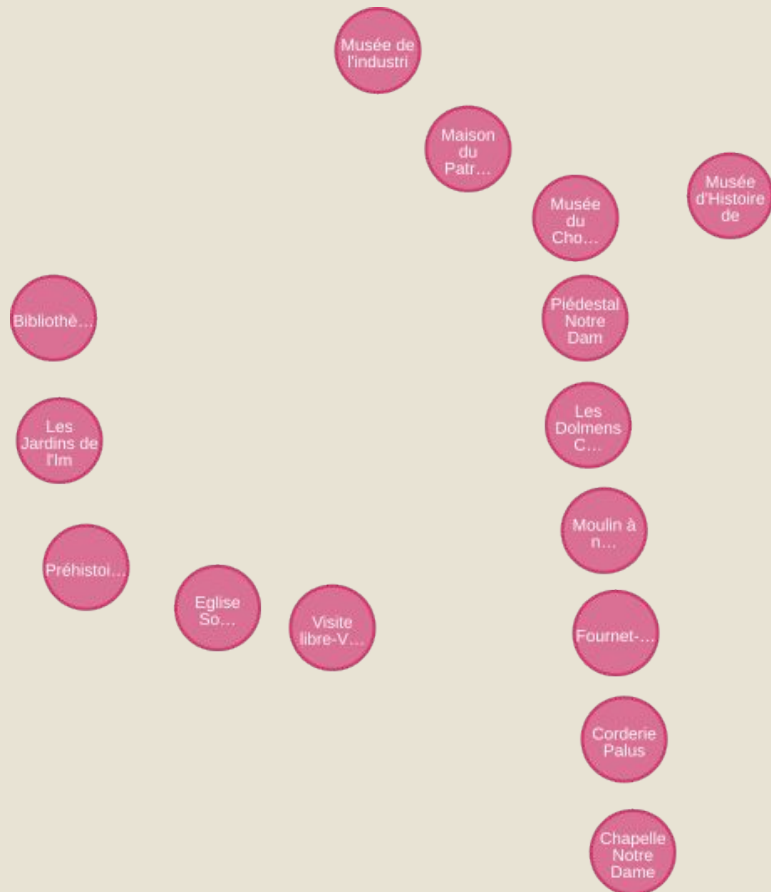


# Stockage dans PostgreSQL





Stockage de l'ensemble des POIs  
dans neo4j pour le calcul  
d'itinéraires thématiques par  
catégorie



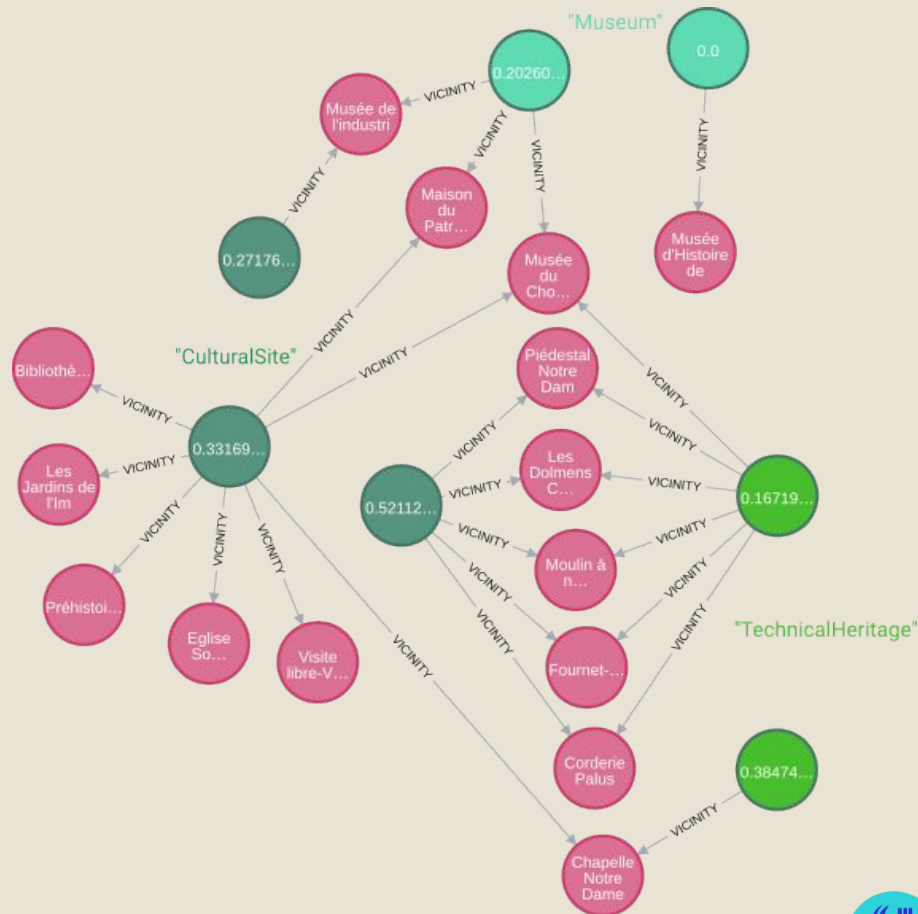


## Clusters et Graphes 2/3



Dans chaque catégorie, les POIs sont regroupés par densité géographique.

- [HDBSCAN](#) / scikit-learn
- VICINITY entre POIs et clusters dans neo4j

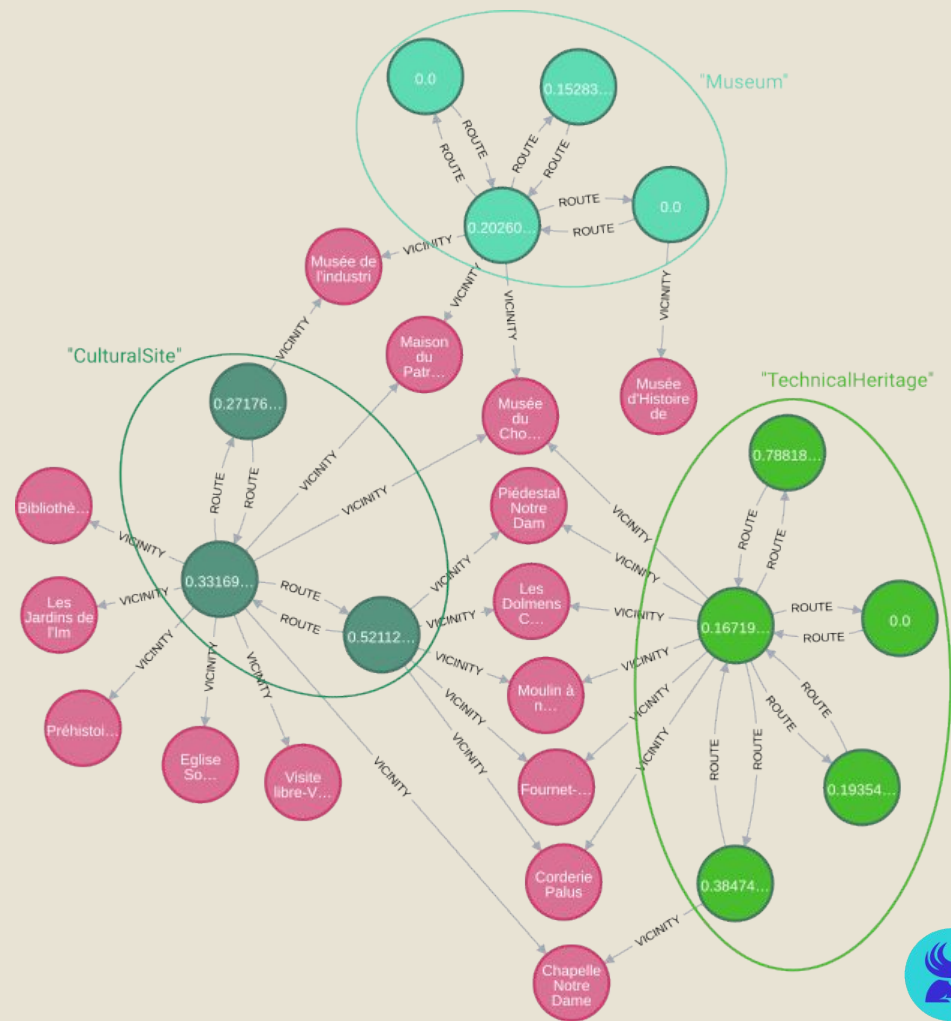


## Clusters et Graphes 3/3



Dans chaque catégorie, génération d'un graphe entre les clusters

- [arbre couvrant de poids minimal](#)
- augmentation du graphe par sélection d'arêtes sur la [triangulation de Delaunay](#)
- ROUTE entre clusters dans neo4j



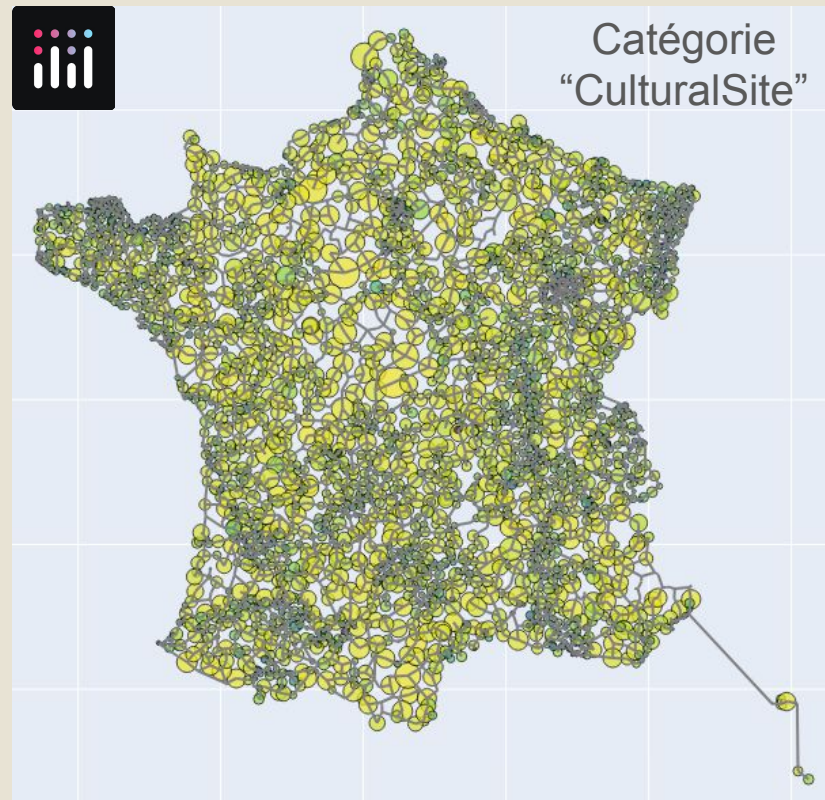
## Qualité des données 1/3



Une application Dash/Plotly fournit une vision globale de la qualité des données issues de DATAtourisme.

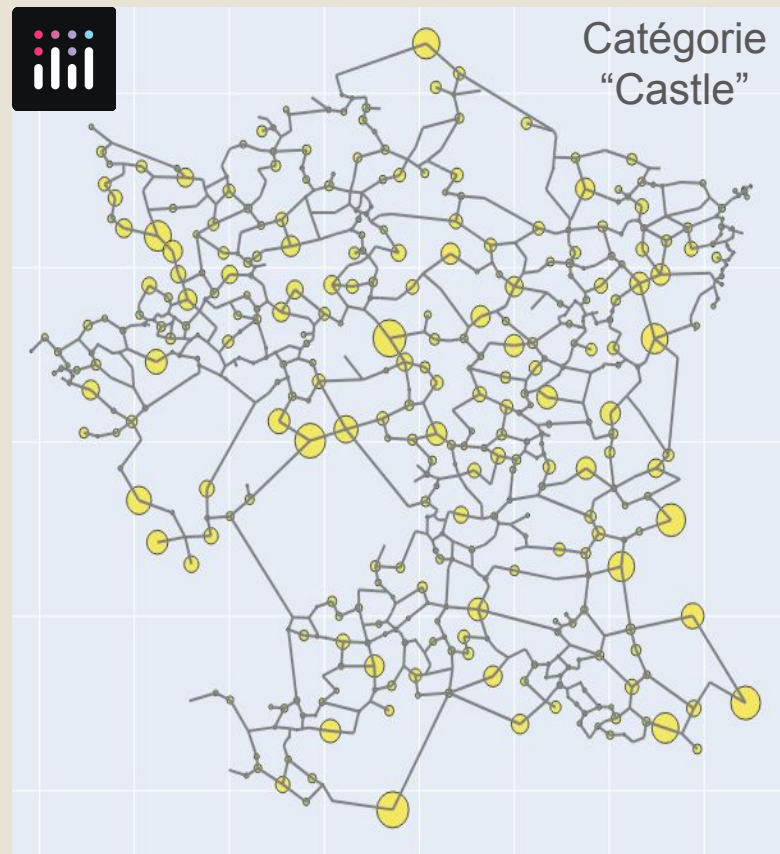
Les catégories les plus utilisées offrent

- Une très bonne couverture du territoire français
- Peu d'intérêt thématique.





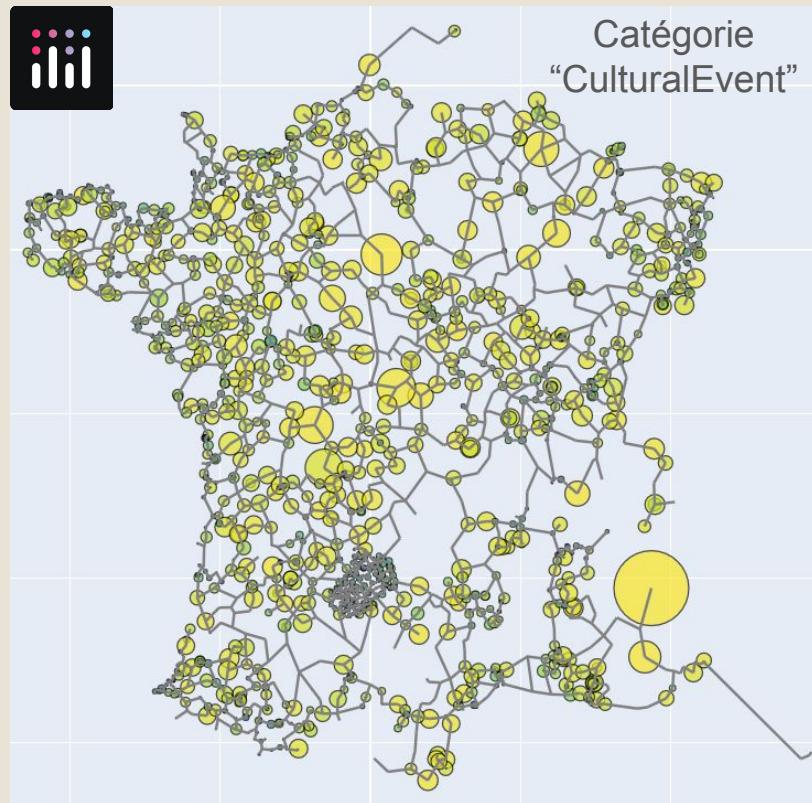
Certaines catégories sont très ciblées mais ne sont pas utilisées de manière uniforme sur l'ensemble du territoire.





Certaines catégories sont sur-utilisées dans certaines régions.

Exemple : CulturalEvent dans le département du Lot.





## Consommation des données



- Plusieurs services peuvent être proposés en exploitant les données disponibles et leur structure
  - Recherche de POI par ville/département/région/catégorie
  - Informations détaillées sur un POI
  - Recherche de ville/département/région par catégorie de POI
  - Voisinage géographique et thématique d'un POI
  - Itinéraire thématique entre POIs
- Exemple d'application : propositions d'itinéraire thématique au cours d'un trajet entre 2 villes

## Itinéraire thématique entre 2 villes

Catégorie :

Ville de départ :

Ville d'arrivée :

### Détails de l'itinéraire

#### Étape 1 (red):

- PNALAR0340000039: ETANG ASSECHE DE MONTADY, 34440 Colombiers (43.3148346,3.1383661)
- PNALAR034V5000N6: NEUF ECLUSES DE FONSERANES, 34500 Béziers (43.3306239,3.1996802)
- PNALAR034V5000P2: LE CANAL DU MIDI, 34500 Béziers (43.332294,3.2038859)
- PNALAR034V51NVDC: ETANG DE LA MATTE, 34710 Lespignan (43.2649522,3.15389015)

#### Étape 2 (blue):

- PNALAR034V51IUKK: ETANG DE CAPESTANG, 34310 Capestang (43.33165,3.043002)

#### Étape 3 (green):

- PNALAR034V50M733: GORGES DE REALS, 34460 Cessenon-sur-Orb (43.436953,3.10152641)

#### Étape 4 (orange):

- PNALAR034V51IUKK: LES BARRES ROCHEUSES DE CAZEDARNES, 34460 Cazedarnes (43.420492,3.0215679)
- PNALAR034V50SP62: FORET DES EUCALYPTUS, 34460 Cessenon-sur-Orb (43.4576356,3.0173091)
- PNALAR034V51IUKB: RESERVE NATURELLE REGIONALE DE COUNIAC, 34460 Cessenon-sur-Orb (43.4688389,3.05826742)

#### Étape 5 (purple):

- PNALAR034V52CAQY: PLAGE DE ROQUEBRUN, 34460 Roquebrun (43.4990168,3.03011203)
- PNALAR034V50LW5I: Tables D'orientation de Berliou, 34360 Berliou (43.4826424,2.97557142)
- PNALAR034V52IXHJ: NAPPE DE MONTPEYROUX A ROQUEBRUN, 34460 Roquebrun (43.50048,3.02861)

#### Étape 6 (darkred):

- PNALAR0340000028: LES VALLEES DE L'ORB ET DU JAUR, 34390 Olargues (43.5192404,2.9924869)
- PNALAR034V52IXHX: NAPPE DE MONTPEYROUX, 34390 Vioussan (43.54069,2.97686)
- PNALAR034V50930P: PIC DE NAUDECH, 34390 Vioussan (43.5366123,2.9513239)

#### Étape 7 (cadetblue):

- PNALAR034V50U6V5: CASCADE LE FREJO, 34390 Olargues (43.5507655,2.9003384)

#### Étape 8 (pink):

- PNALAR034V52HNVJ: Lac de vésoles, 34330 Fraisse-sur-Agout (43.5578232,2.796958)
- PNALAR034V5095F0: MONTS DU SOMAIL, 34220 Saint-Pons-de-Thomières (43.5605382,2.8247423)
- PNALAR0340000032: LE LAC ET SAUT DE VESOLES, 34390 Prémian (43.5531513,2.7941751)

#### Étape 9 (darkblue):

- LOILAR034V50XKY9: GROTTE DE LA FILEUSE DE VERRE, 34220 Courniou (43.4738918,2.7132587)
- PNALAR0340000013: MASSIF DU CAROUX-ESPINOUSE, 34220 Saint-Pons-de-Thomières (43.488666,2.75794)
- PNALAR034V52C2OU: SOURCE DU JAUR, 34220 Saint-Pons-de-Thomières (43.4871803,2.75754667)

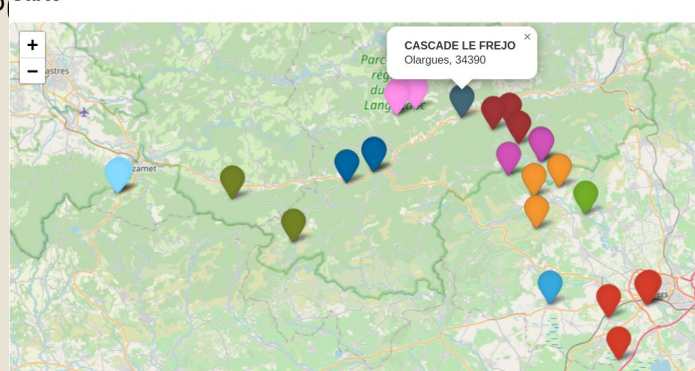
#### Étape 10 (darkgreen):

- PNALAR034V52C2SZ: SOURCE DE LA CESSÉ, 34210 Ferrals-Les-Montagnes (43.4037434,2.62763261)
- 7027835: Lac d'Albine, 81240 Albine (43.454928,2.52886)

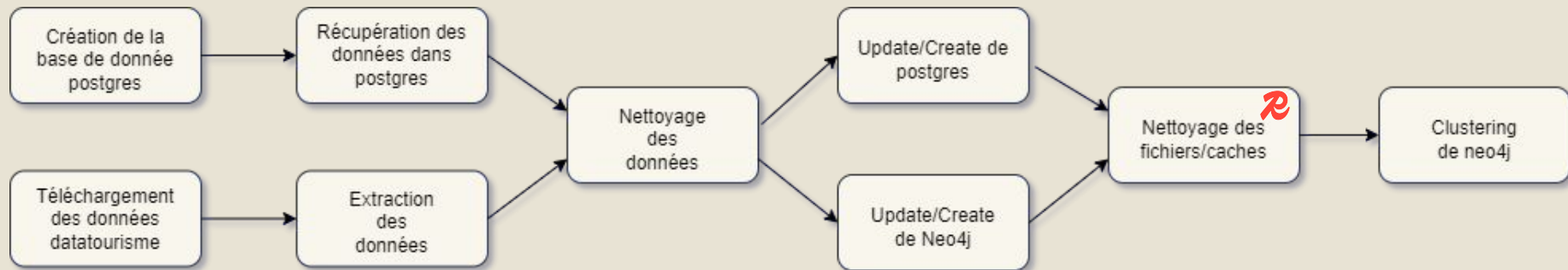
#### Étape 11 (lightblue):

- 844917: Base de Loisirs du Lac des Montagnès, 81200 Mazamet (43.462109,2.341805)
- 4675389: Fautueil de mise à l'eau PMR, 81200 Mazamet (43.463401,2.345725)

### Carte



# Orchestration



# Monitoring

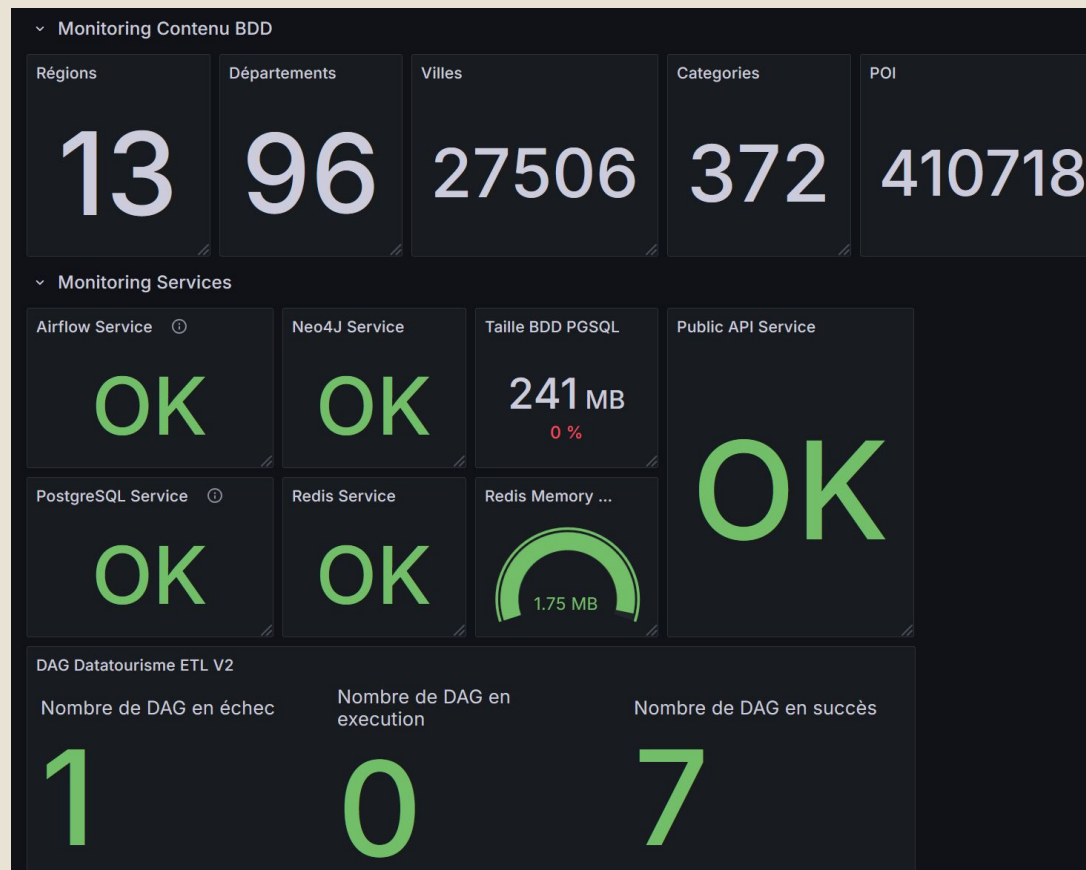


Prometheus pour aller scrapper les metrics des services suivants à l'aide d'exporters :

- PostgreSQL
- Redis
- Airflow
- API Publique (via blackbox)
- Neo4J (via blackbox)

On vient monitorer 2 types de métriques :

- Contenu de la BDD
- Etat des services



pgsql\_exporter:9187  
airflow\_exporter:9112  
blackbox:9115  
redis\_exporter:9121

:9090



:3000





## Améliorations / évolutions possibles



- Possibilités de filtrage selon plusieurs critères, par exemple :
  - Filtrage par rating
  - Distance maximum à parcourir
  - Temps maximum du circuit de visite
  - Dépense maximum du circuit de visite
- Sécurisation de l'API:
  - TLS + Authentification
  - Rate limiting sur les endpoints "sensibles" (compute / authentification) (pour éviter un denial of service)

### Contraintes Techniques / Coût :

- Acquisition et extraction de données :
  - Risque de bannissement d'IP suite à un scrapping trop intensif (Exemple : scrapping du frontend de Tripadvisor)
  - Limitation de la fréquence du refresh des données sur [Datatourisme](#) (1 fois par 24h)
- Coûts d'usage d'API externes pour enrichir le jeu de données : (rating, commentaires...)
  - Google Places : 32\$ pour 1000 requêtes / Sachant qu'une recherche peut utiliser **plusieurs** appels avec la pagination.



# Démonstration

