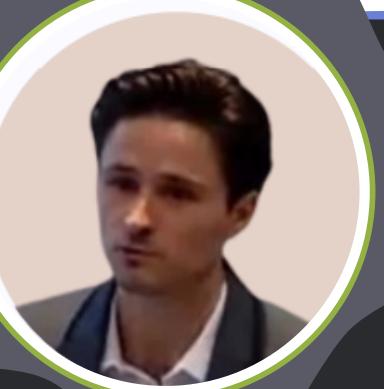


Soutenance

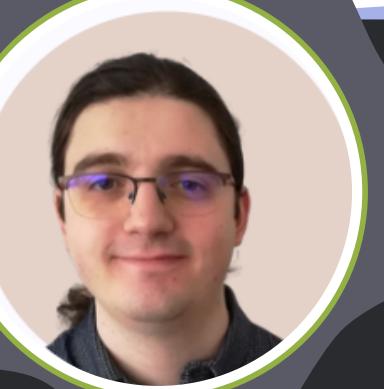
Architecte BIG DATA et science des données



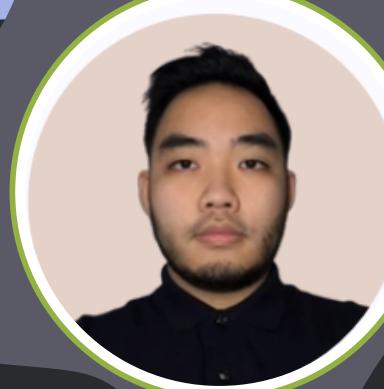
⚠ Afin de regarder les vidéos de démonstration, merci de vous rendre sur Canva en [cliquant ici](#)



**GUILLAUME
BOTTAZ-BOSSON**



**THOMAS
LAHAYE**



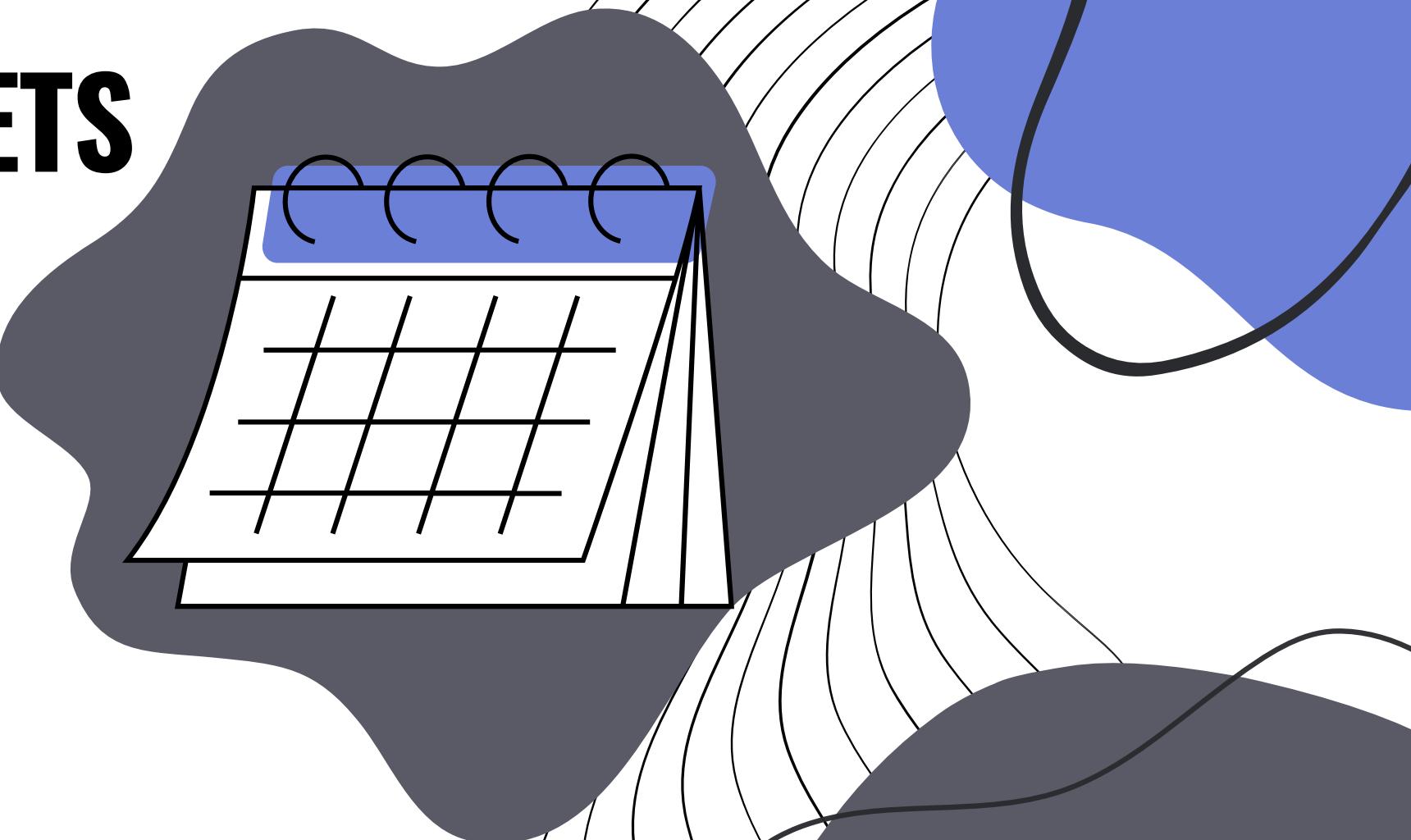
**JOHAN
LO**

PRÉSENTATION DES PROJETS

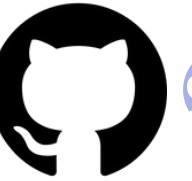
Projet 1
Projet 2
Projet 3

Scrapper
Infrastructure big data
Machine learning

- ◆ Contexte & Objectifs
- ◆ Architecture du projet
- ◆ Fonctionnement du programme



PROJET 1 : SCRAPPER DE SITE WEB



PROJET 1 : SCRAPPER DE SITE

◆ Contexte & Objectifs

Présentation

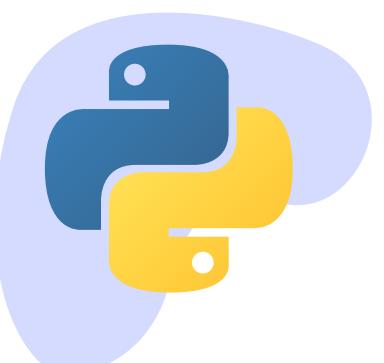
Contexte

Quels sont les objectifs du projet ?
Qu'est ce qu'un scrapper ?



 mongoDB

 HTML
5

 Python

PROJET 1 : SCRAPPER DE SITE

♦ Contexte & Objectifs

Méthodologie

URL de base

Mise en place de filtre pour la qualité de la donnée

Définition d'un scope

Etape 1

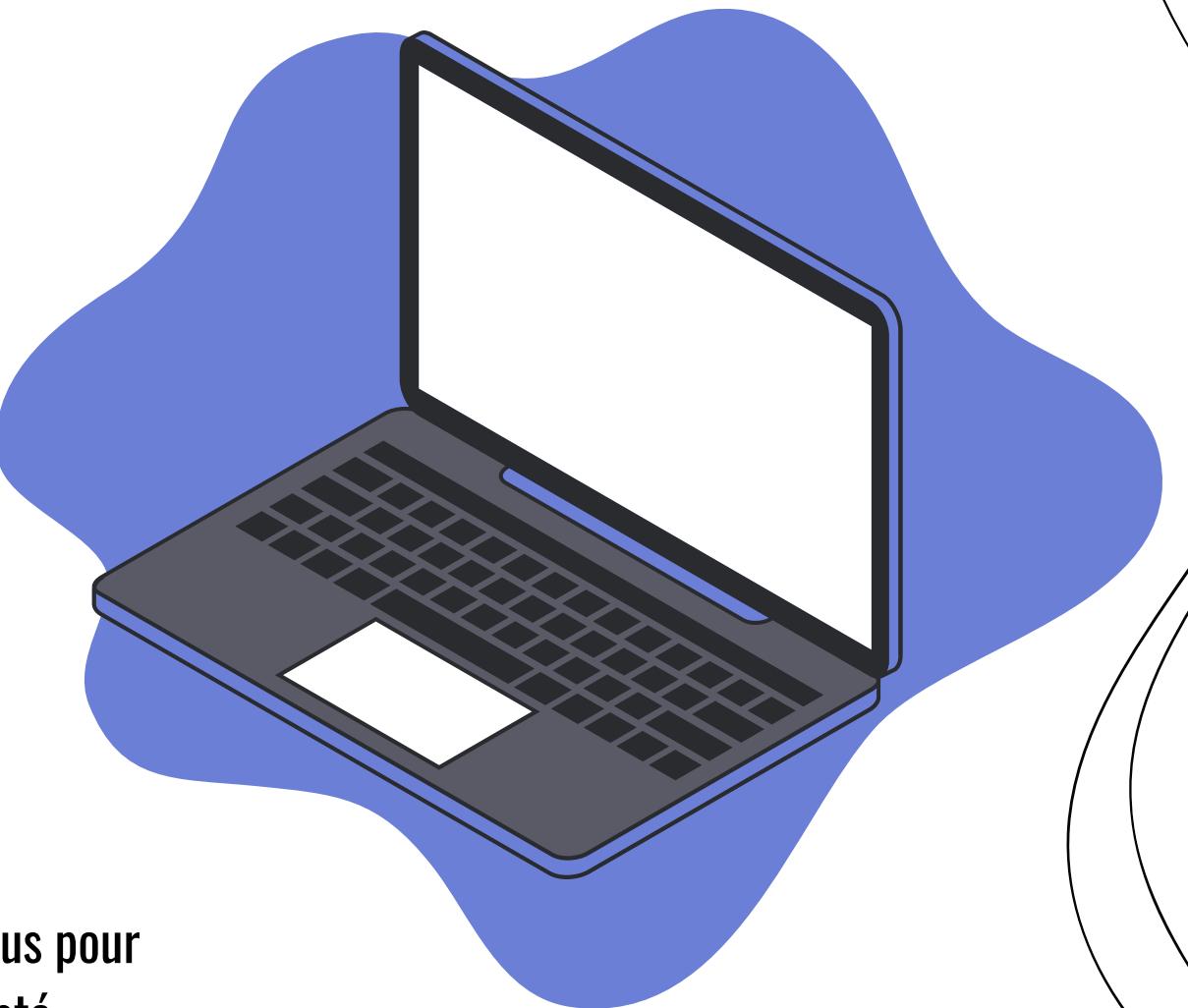
scraping des éléments du site et ajout dans une liste

Etape 2

analyse des liens pour prendre uniquement les liens internes

Etape 3

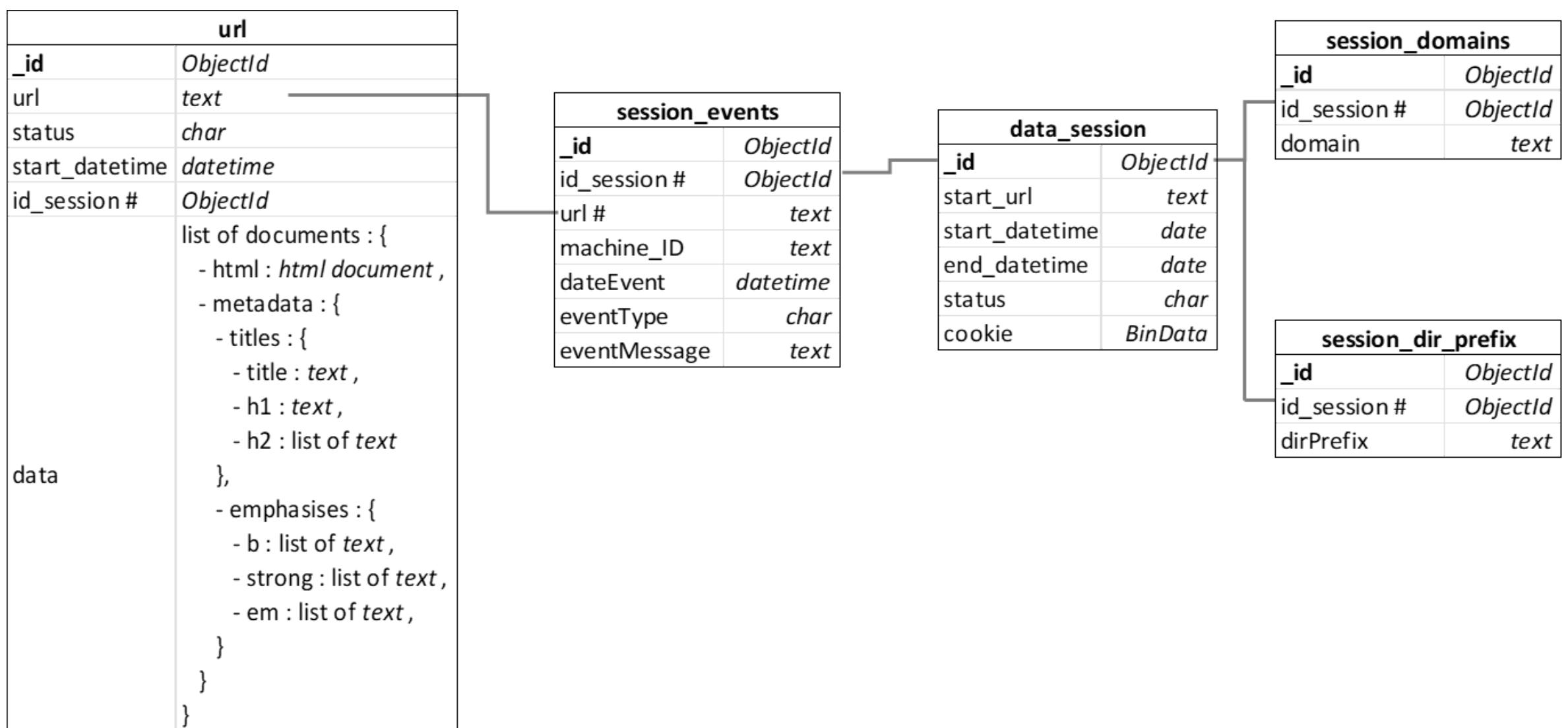
répéter le processus pour chaque élément listé.



PROJET 1 : SCRAPPER DE SITE

◆ Fonctionnement

Architecture de la base de données



PROJET 1 : SCRAPPER DE SITE

♦ Fonctionnement

Exemple de documents

Data_session

```
_id: ObjectId('64a7ce3fad9cec0f3a534aee')
start_url: "https://fr.wikipedia.org/wiki/Garnier_de_Rochefort"
start_datetime: 2023-07-07T10:35:11.944+00:00
status: "done"
cookies: "<RequestsCookieJar[<Cookie WMF-Last-Access=07-Jul-2023 for fr.wikipedia.org>]"
```

session_events

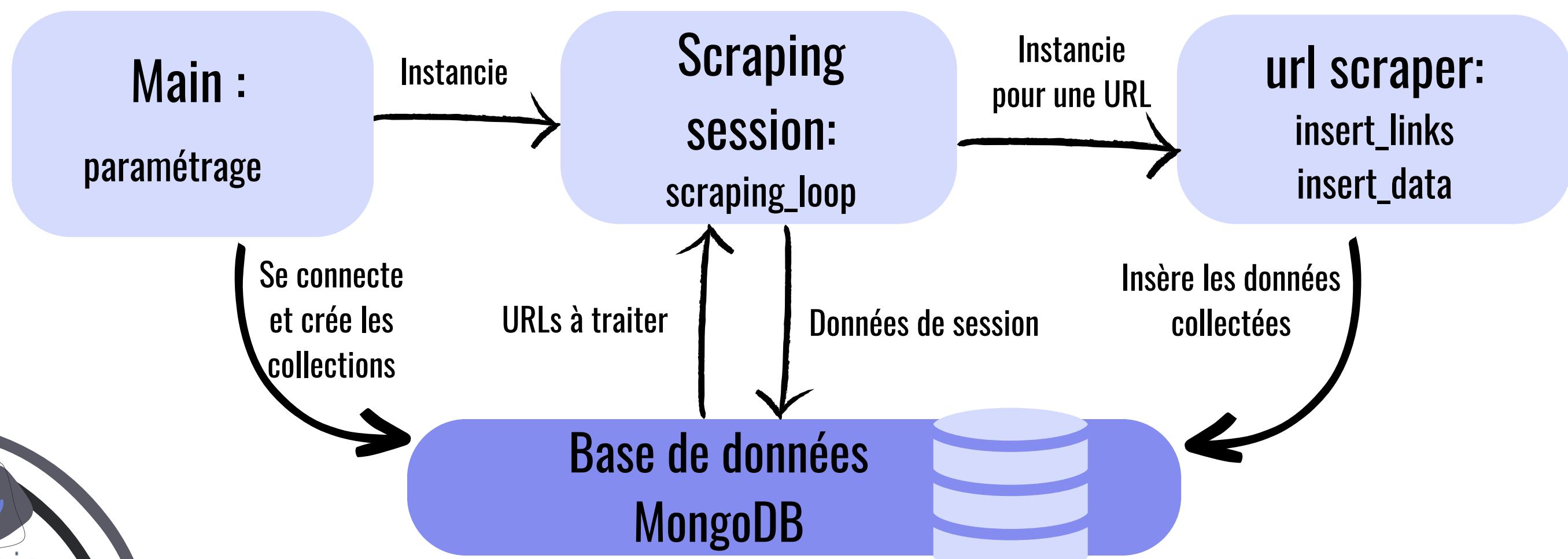
```
_id: ObjectId('64a7ce3fad9cec0f3a534af0')
id_session: ObjectId('64a7ce3fad9cec0f3a534aee')
machine_ID: "DESKTOP-T7CHBTW"
datetime: 2023-07-07T10:35:11.957+00:00
event_type: "Session starting"

dateEvent: 2023-07-07T10:35:11.993+00:00
eventType: "launch url scraping"
eventMessage: "launch scraping on https://fr.wikipedia.org/wiki/Garnier_de_Rochefort"

eventType: "url scraping abortion"
eventMessage: "Traceback (most recent call last):
  File "C:\Users\digidi\PycharmProj..."
```

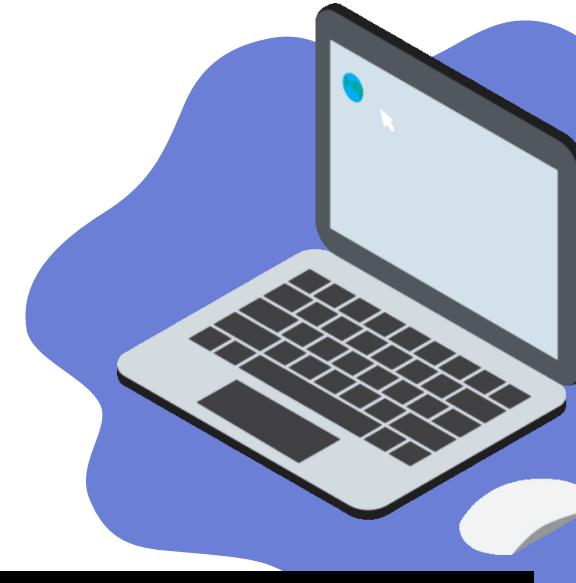
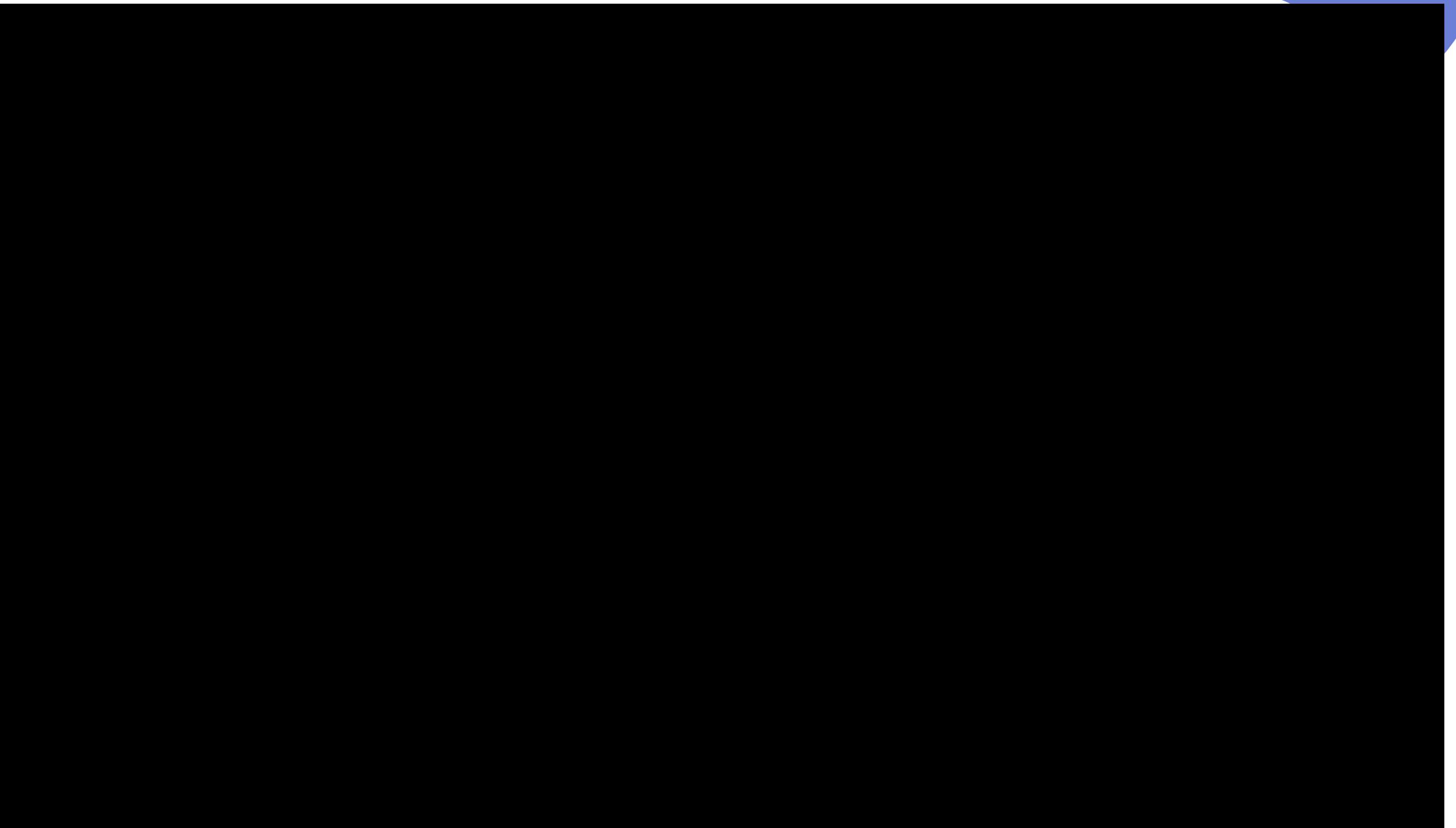
PROJET 1 : SCRAPPER DE SITE

◆ Démonstration Fonctionnement

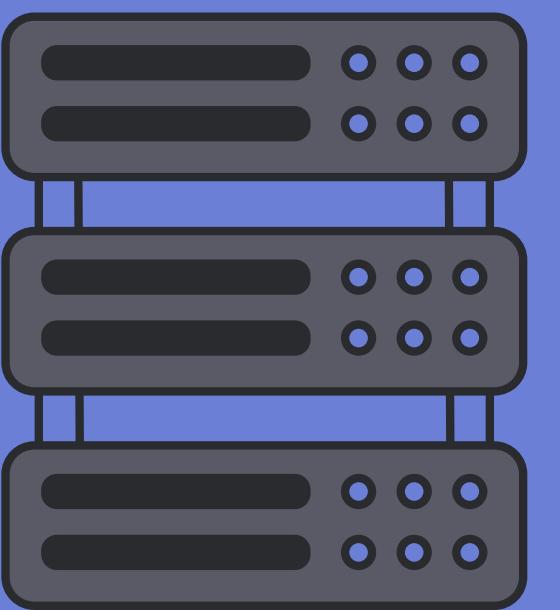


PROJET 1 : SCRAPPER DE SITE

◆ Démonstration



PROJET 2 : INFRASTRUCTURE BIG DATA



PROJET 2 : INFRASTRUCTURE BIG DATA

◆ Problématique - Activité du client

Commercialisation de produits laitiers



- # Fromagerie industrielle

Département de la Mayenne (53)



Boutique de fidélisation

- Des points de fidélité à récolter sur les produits laitiers
 - Échangeables contre des objets

-> Stockage des données de gestion des commandes dans un Data Warehouse



PROJET 2 : INFRASTRUCTURE BIG DATA

◆ Problématique - Besoin exprimé

Extraction d'indicateurs et rapports statistiques



Deux lots :

- Lot 1 : indicateurs sur
 - commandes du département Mayenne (53) passées après 2010 avec une référence objet commandée plus de 5 fois
 - commandes des 10 clients les plus fidèle depuis 2008
 - croissance par objet pour les départements du Maine et Loir (49), de la Mayenne (53) et de la Sarthe (72)
- Lot 2 : indicateurs sur les 100 meilleures commandes

Infrastructures



Lot 3 :

- Mise en place d'une base NoSQL ainsi que d'un moteur de recherche pour interroger le Data Warehouse
- Pour répondre aux lots 1 et 2.

PROJET 2 : INFRASTRUCTURE BIG DATA

♦Notre démarche

Comprendre les données

Faire des choix

Ébauche à la main
(tables et graphiques)

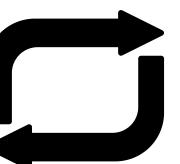
Questionnement client

**COMPRENDRE LE BESOIN CLIENT
(LOTS 1 & 2)**

**MISE EN OEUVRE EN LOCAL
(LOTS 1 & 2)**

Mise en place base NoSQL
Liaison Power BI

**FONCTIONNEMENT SUR INFRASTRUCTURE DISTRIBUÉE HADOOP
(LOT 3)**



PROJET 2 : INFRASTRUCTURE BIG DATA

♦Les Données

135 000 lignes*

*détail d'une commande passée par un client pour un objet

25 colonnes

5 sans valeur

(6) données client

(10) données commande

(7) données objet commandé

(2) données conditionnement



Hypothèses :

- Les duplications de lignes sont fiables
- Les offres promotionnelles sont saisies lors de leur commande d'utilisation



PROJET 2 : INFRASTRUCTURE BIG DATA

♦Les Données

Qualité des données

Etat des lieux : Data Warehouse

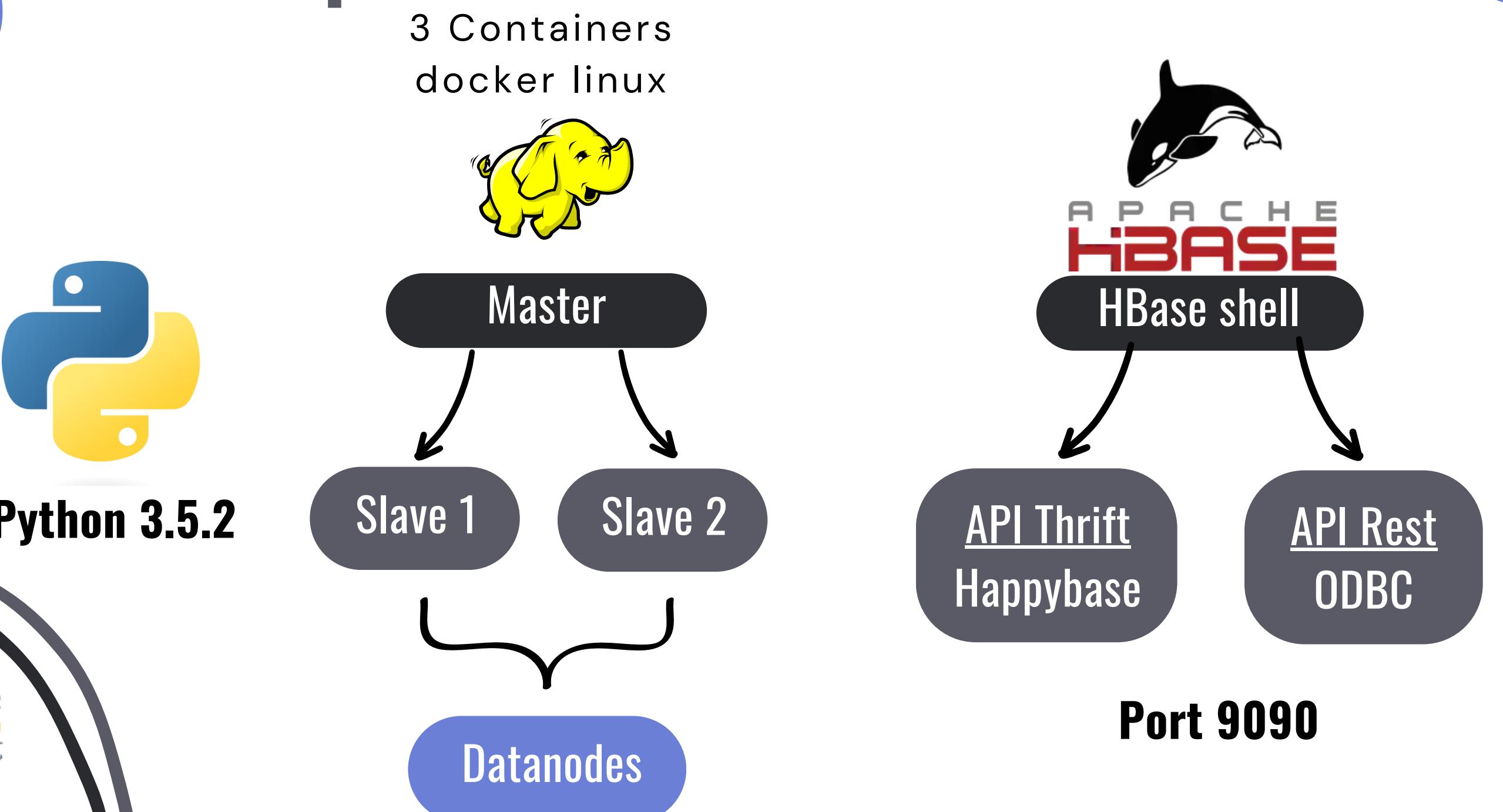
ETL

L'importance d'une donnée de qualité



PROJET 2 : INFRASTRUCTURE BIG DATA

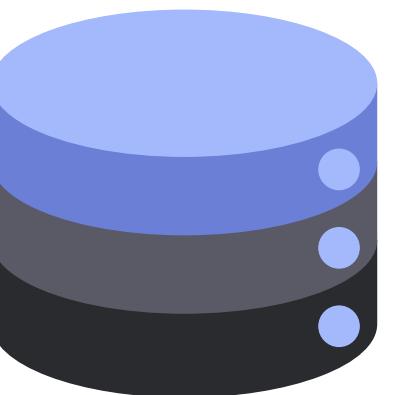
◆ Présentation de l'architecture **Hadoop**



PROJET 2 : INFRASTRUCTURE BIG DATA

◆ Présentation de l'architecture
Job Map-Reduce

Données client

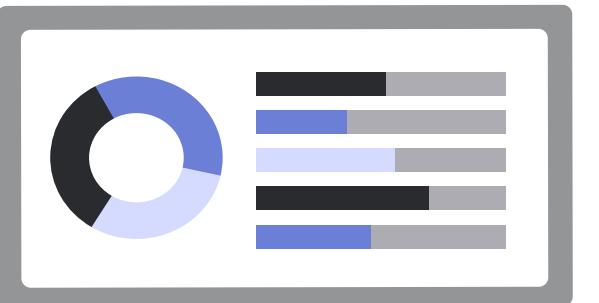


Mapper

Données triées & uniformisées



Reducer



Données regroupées

PROJET 2 : INFRASTRUCTURE BIG DATA

♦Présentation de l'architecture **Scripts de commandes**

Fichier .bat:

- Import des fichiers .csv, .sh et .py dans le container “master”
- Démarrage des services sur slave 1 et 2
- Exécution du fichier.sh
- Lancement du job map-reduce (.jar)

Fichier .sh:

- Démarrage Hadoop
- Démarrage HBase
- Ouverture de l'API Thrift

PROJET 2 : INFRASTRUCTURE BIG DATA

◆ Focus : Lot 1.3

Des données brutes...

Le client souhaite : une courbe de croissance par objet vendu sur 3 départements

```
dataw_fro03.csv
1 , "prenomcli", "cpcli", "villecli", "codcde", "datcde", "timbrecli", "timbrecde", "Nbcolis", "cheqcli", "barchive", "bstock", "codobj", "qte", "Colis", "libobj", "Tailleobj",
2 .el", "14540", "BOURGUEBUS", "478", "2004-10-22 00:00:00", "5", "4.8", "1", "NULL", "1", "1", "38", "2", "1", "Polo", "XL", "230", "60", "0", "Carton Tete de menagere", "0", "0"
3 .el", "14540", "BOURGUEBUS", "478", "2004-10-22 00:00:00", "5", "4.8", "1", "NULL", "1", "1", "30", "2", "1", "T-shirt Blanc", "L", "170", "60", "0", "Carton Tete de menagere", "0", "0"
4 .el", "14540", "BOURGUEBUS", "478", "2004-10-22 00:00:00", "5", "4.8", "1", "NULL", "1", "1", "45", "1", "1", "Montre", "Homme", "30", "150", "0", "Carton Tete de menagere", "0", "0"
5 iony", "35400", "SAINT MALO", "21239", "2006-10-03 00:00:00", "0", "3.9", "1", "1.45", "1", "1", "31", "1", "1", "T-shirt Blanc", "XL", "180", "60", "0", "Distingo 500 g", "34", "0"
```

Mapper:

Reducer:



PROJET 2 : INFRASTRUCTURE BIG DATA

◆ Focus : Lot 1.3
... au mapper ...

Récuperer
codobj, depcli, year,
codcde, libobj, qte

```
for line in sys.stdin:  
  
    line = line.strip()  
  
    datafro = line.split(',')  
    if len(datafro) < 24 :  
        print('AN\t-%s' % line)  
    else:  
        for n in range(len(datafro)):  
            datafro[n] = datafro[n].strip('')  
  
        codcli = datafro[0]  
        codecde = datafro[6]  
        villecli = datafro[5]  
        cpcli = datafro[4][:2]  
        nbcoli = datafro[10]  
        timbrecde = datafro[9]  
        date = datafro[7].split('-')[0]  
        qte = datafro[15]  
        point = datafro[20]  
        try:  
            year = int(date)
```

Critère:
Département
72, 53, 49

PROJET 2 : INFRASTRUCTURE BIG DATA

◆ Focus : Lot 1.3
... au reducer ...

```
if current_dep and current_dep == depcli and current_year == year and current_obj == codobj:  
    # On vérifie que la commande est la même qu'à la ligne précédente pour sommer la quantité  
    if current_cde == codcde:  
        current_qte = qte  
    else:  
        current_nbcde += 1  
        current_cde = codcde  
        current_qte += qte  
    else:  
        # On vérifie que current_dep est bien définie pour afficher le résultat  
        if current_dep:  
            # ajout de la ligne au dataframe  
            row = {'objet':current_libobj, 'departement':current_dep, 'annee':current_year,  
                   'nombre_commandes':current_nbcde, 'quantite':current_qte}  
            data = pd.concat([data, pd.DataFrame([row])], ignore_index=True)
```

Calcul les totaux
/produit/an/département

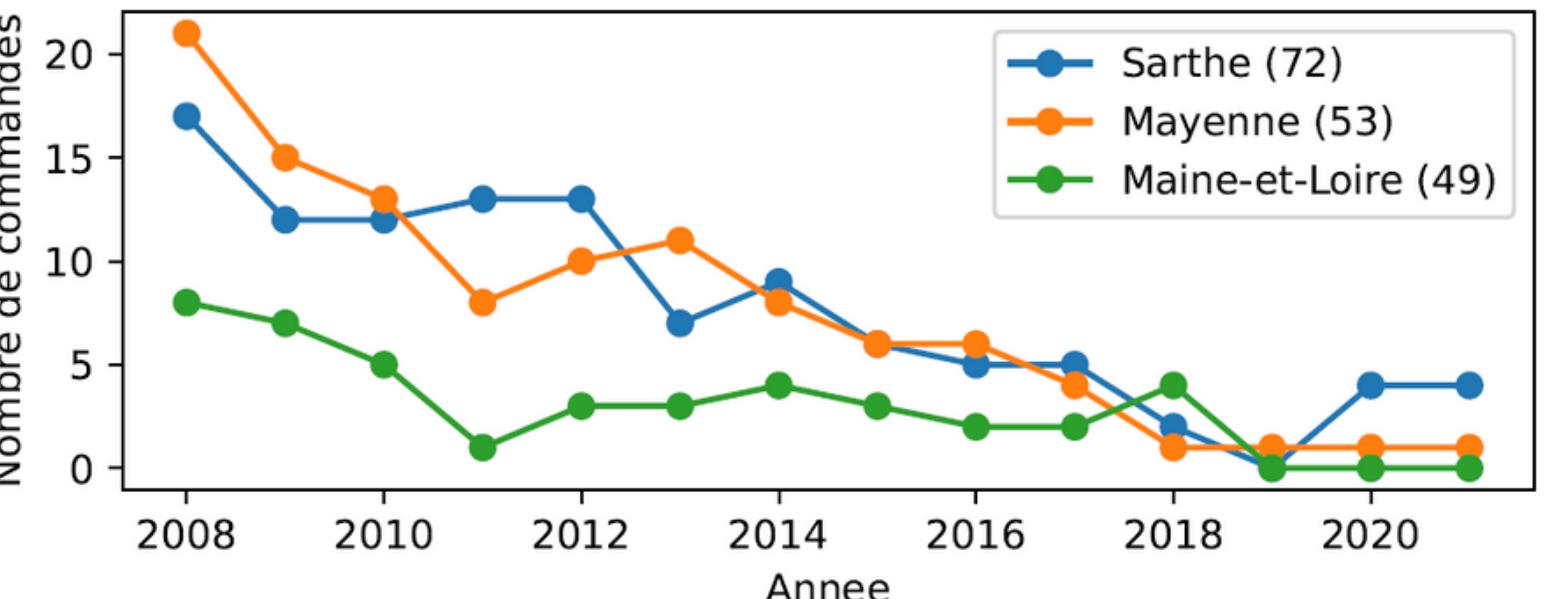
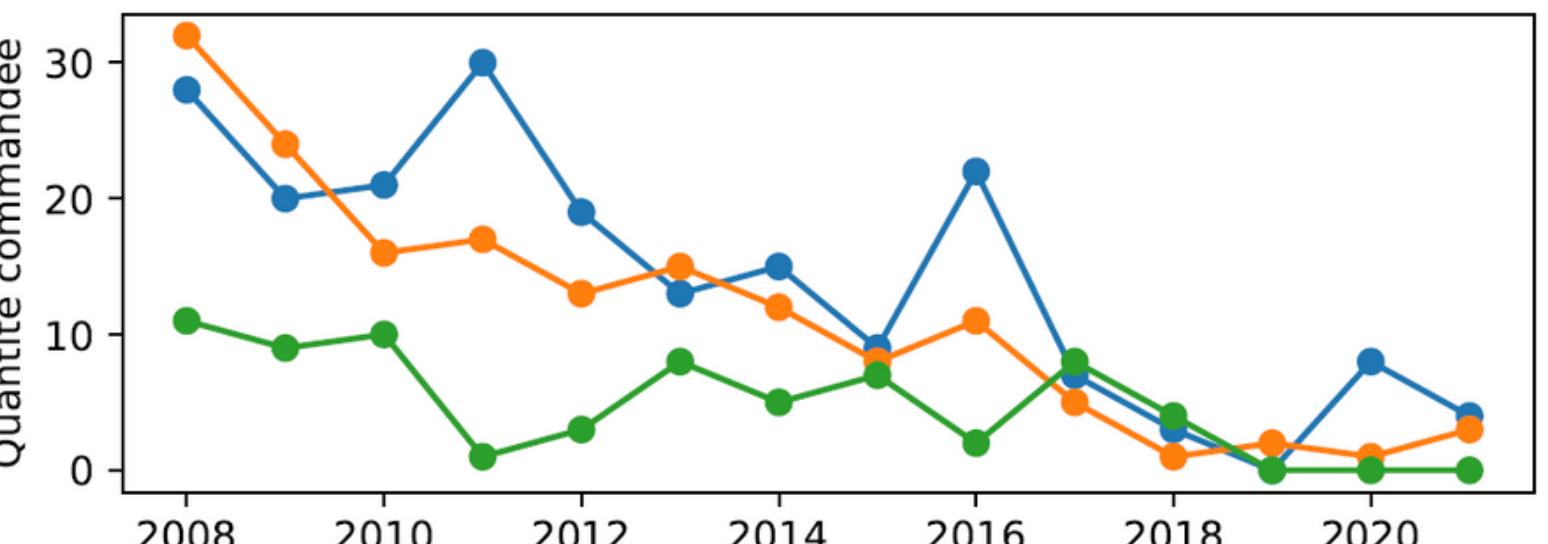


PROJET 2 : INFRASTRUCTURE BIG DATA

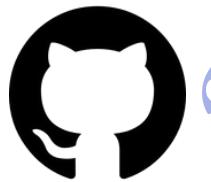
♦ Focus : Lot 1.3
... au résultat

Rapport en .pdf
Gestion des zéros dans le tracé
des courbes

T-shirt Blanc (XL)



PROJET 3 : MACHINE LEARNING



Source: https://github.com/tholahaye/Projet_scrapper

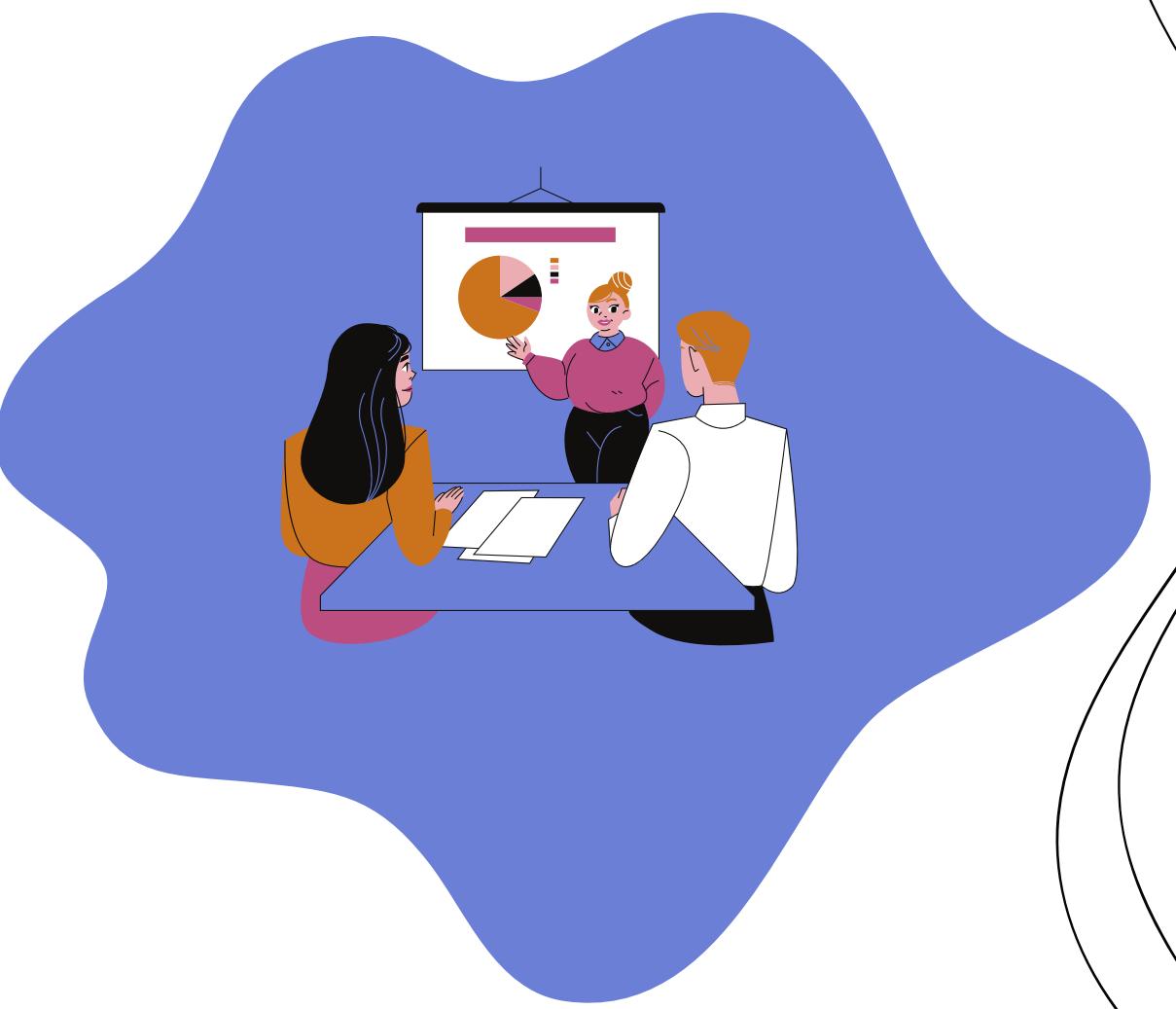
PROJET 3 : MACHINE LEARNING

◆ Contexte & Objectifs

Présentation

Mise en place d'une application Web de machine learning

- Base de données PostgreSQL
- Apprentissage supervisé : Régression & Classification
- Deux modèles implémentés pour chaque cas d'usage
 - Optimisation d'hyper paramètre par validation croisée



PROJET 3 : MACHINE LEARNING

◆ Contexte & Objectifs Les datasets

NOM	TYPE
fixed_acidity	float
volatile_acidity	float
citric_acid	float
df.dtypes	float
chlorides	float
free_sulfur_dioxide	float
total_sulfur_dioxide	float
density	float
ph	float
sulphates	float
Alcohol	float
Target	int

NOM	TYPE
nombre_grossesse	float
concentration_glucose	float
pression_arterielle	float
epaisseur_pli_cutane	float
insuline_serique	float
imc	float
fct_genealogique_diabete	float
age	float
Target	str

PROJET 3 : MACHINE LEARNING

◆ Présentation de l'architecture

Général

AppWeb

Interface web, collecte des paramètres et lancement des traitements

Preprocessing

Prétraitement des données

Machine Learning

Instanciation des modèles,
apprentissage, évaluation des perf.
sélection d'hyper-paramètres

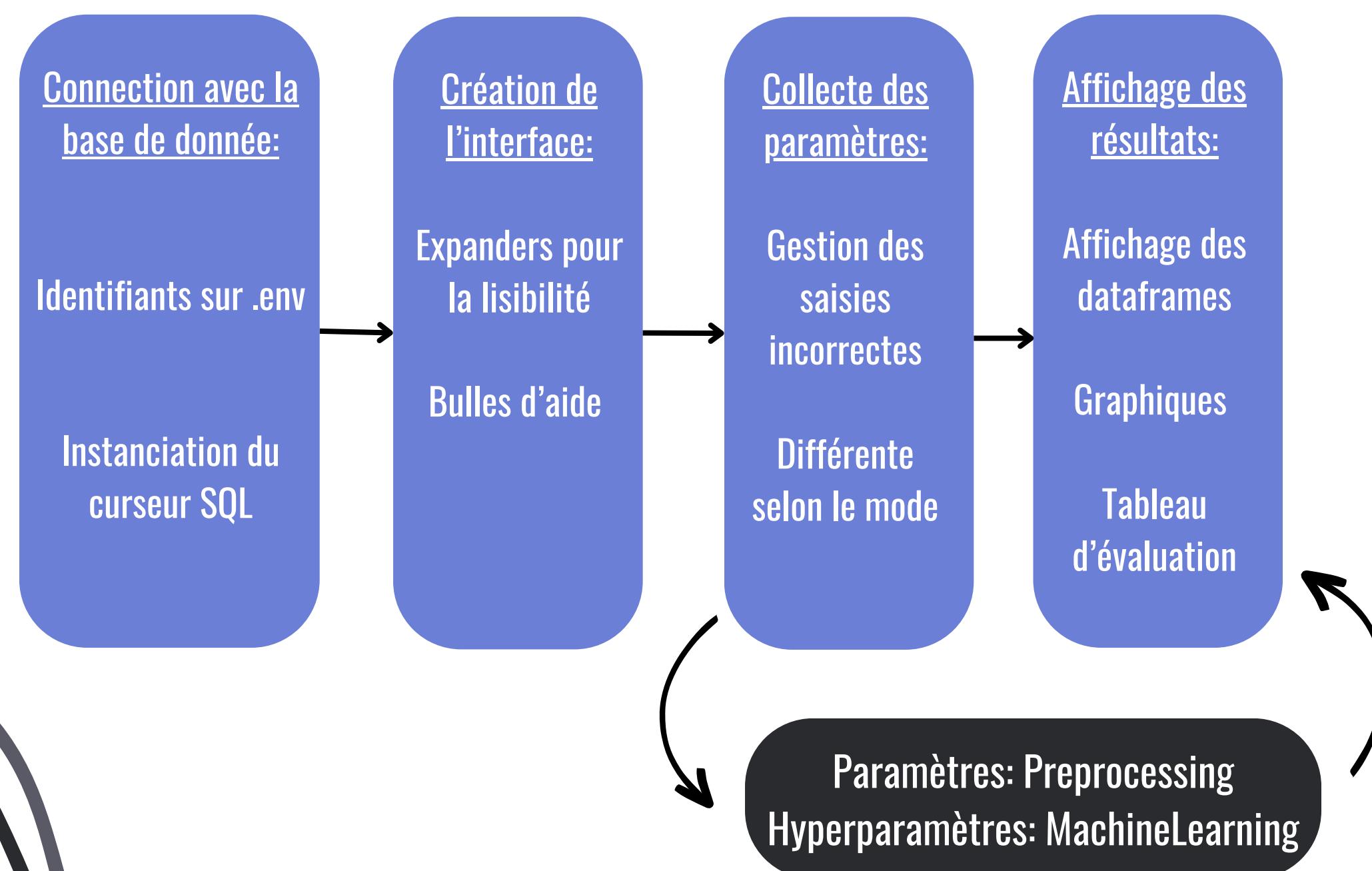
Constantes

Métadonnées des modèles,
Constantes



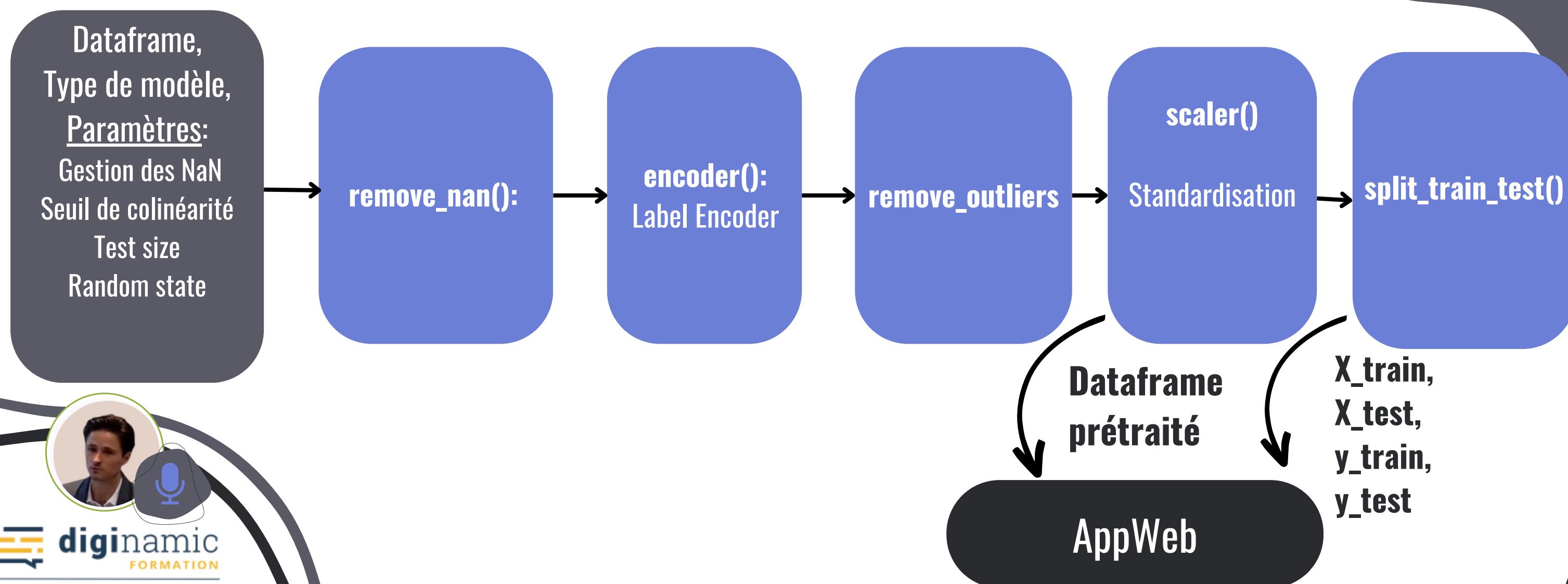
PROJET 3 : MACHINE LEARNING

◆ Présentation de l'architecture AppWeb



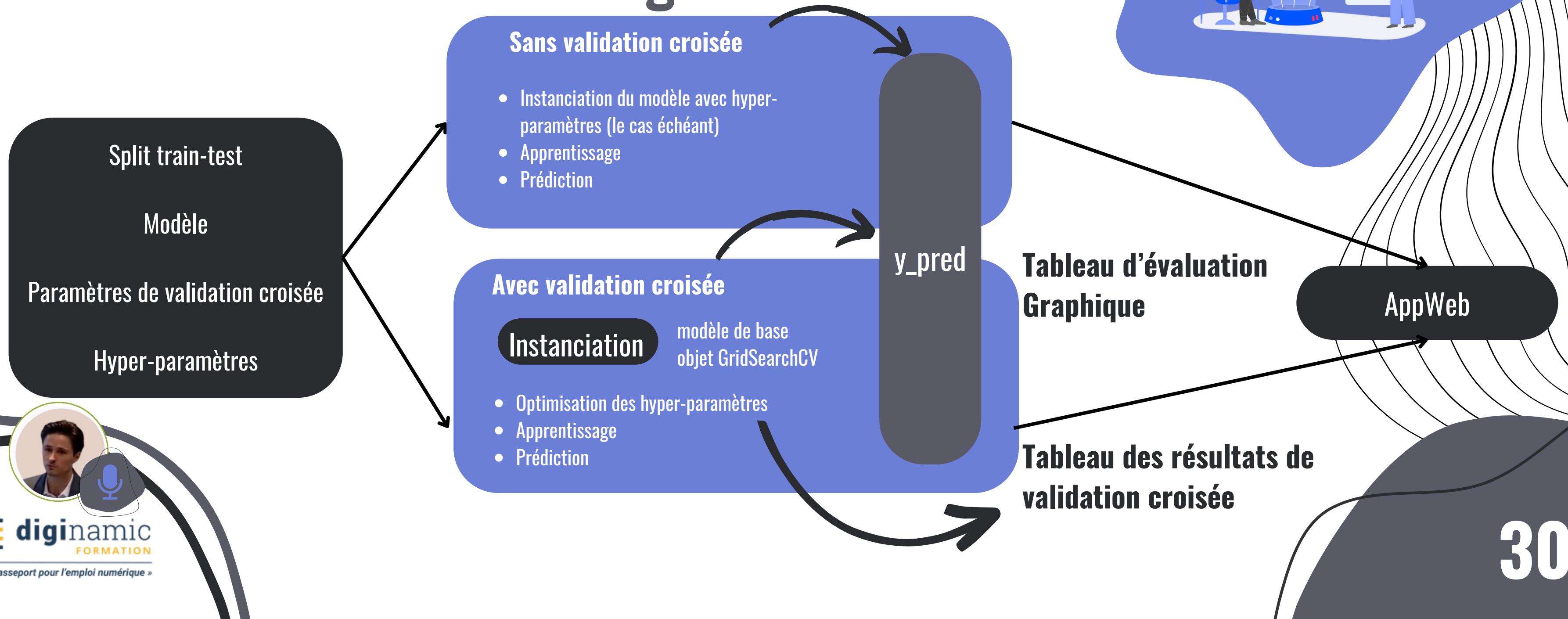
PROJET 3 : MACHINE LEARNING

◆ Présentation de l'architecture Preprocessing



PROJET 3 : MACHINE LEARNING

◆ Présentation de l'architecture Machine Learning

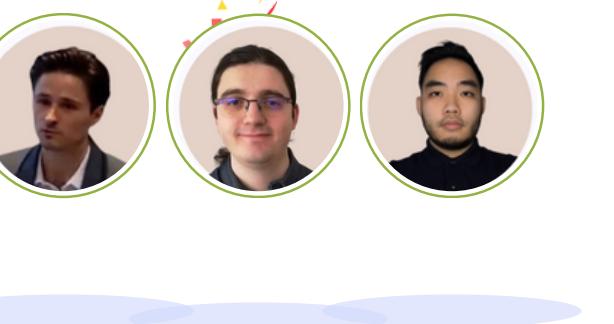


PROJET 3 : MACHINE LEARNING

◆ Démonstration

The screenshot shows a web browser window titled "ML Playground" with the URL "localhost:8501". The interface is dark-themed and features a sidebar on the left labeled "Choose the parameters:" with dropdown menus for "Dataset" (set to "diabete_inde"), "Preprocessing parameters", "Model" (set to "Decision_Tree"), and "Hyperparameters". A checkbox for "Compare several parameters configurations with Cross-validation" is checked. The main area is titled "Machine learning:" and contains three dropdown menus: "Original dataframe", "Processed dataframe", and "Evaluation". At the bottom right of the main area, it says "Made with Streamlit".

CONCLUSION: FIN DE SOUTENANCE **MERCI**



ANNEXE(S)

PROJET 2 : INFRASTRUCTURE BIG DATA

Présentation de l'architecture Scripts de commandes

lot3.bat

```
docker cp lot3_1.sh hadoop-master:/root/  
docker cp lot3_2.sh hadoop-master:/root/  
docker exec hadoop-slave1 /bin/bash -c './service_slv.sh'  
docker exec hadoop-slave2 /bin/bash -c './service_slv.sh'  
docker exec hadoop-master /bin/bash -c './lot3_1.sh'  
python lot3.py dataw_fro03.csv  
docker exec hadoop-master /bin/bash -c './lot3_2.sh'
```

lot3_1.sh

```
./start-hadoop.sh  
start-hbase.sh  
hbase-daemon.sh start thrift
```

lot3_2.sh

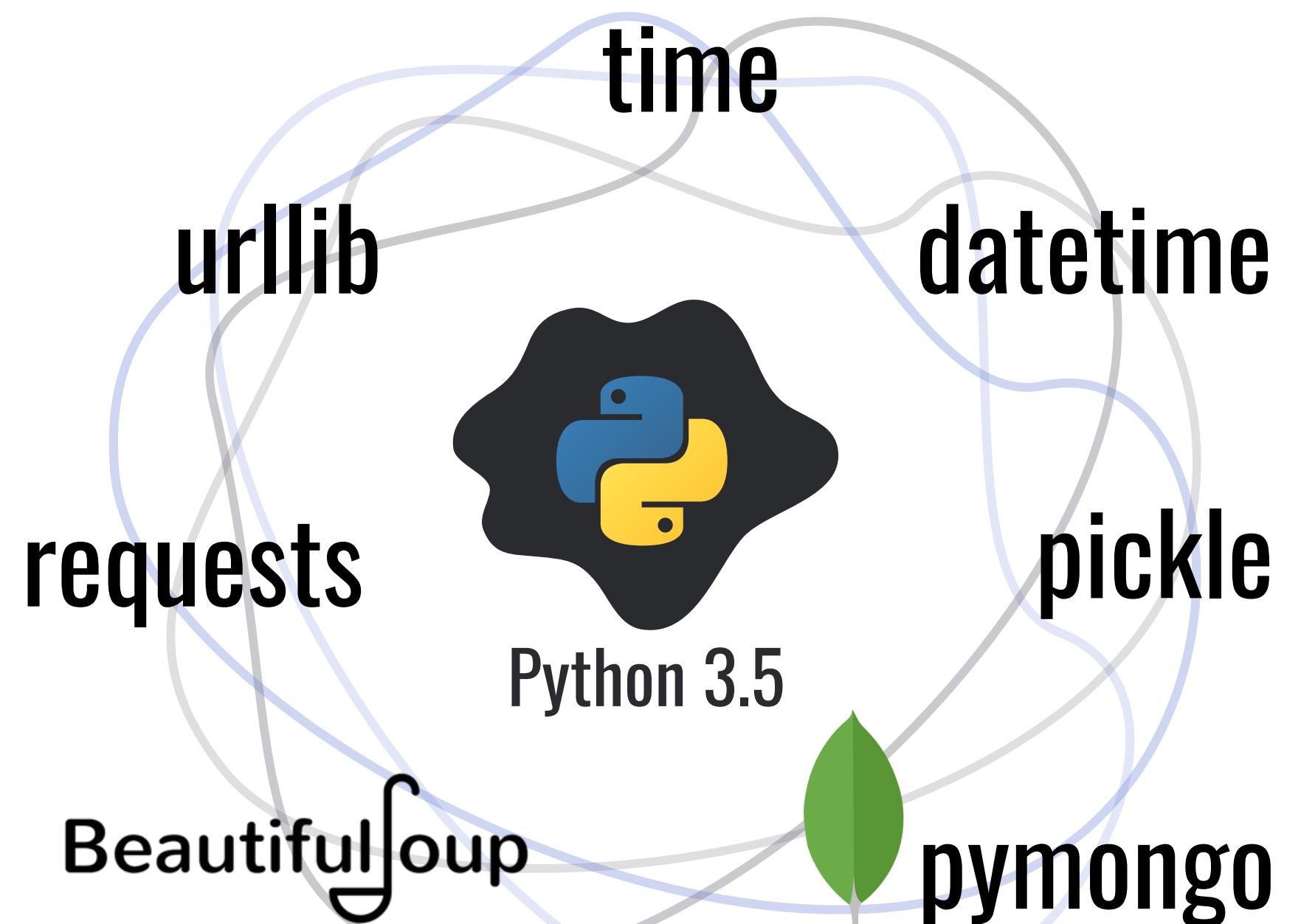
```
hbase-daemon.sh stop thrift  
hbase-daemon.sh start rest -p 9090
```

PROJET 1 : SCRAPPER DE SITE



Mise en place de l'environnement

Librairies



PROJET 2 : INFRASTRUCTURE BIG DATA

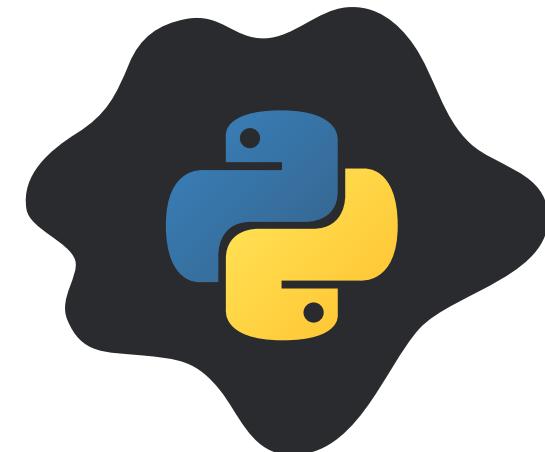


Mise en place de l'environnement

Librairies

OpenPyXL

HappyBase



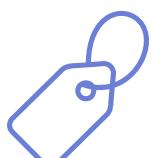
Python 3.5



pandas

matplotlib

PROJET 3 : MACHINE LEARNING



Mise en place de l'environnement

Librairies

