



BIG DATA PROJECT

Soutenance Intermédiaire

**BERGAMIN Maximilien
JAULGEY Thomas
REYNARD Thibaut**



RECAPITULATIF DU SUJET

01



GESTION DE PROJET

02

SOMMAIRE

03

NOTRE AVANCEMENT

04

LA SUITE DU PROJET



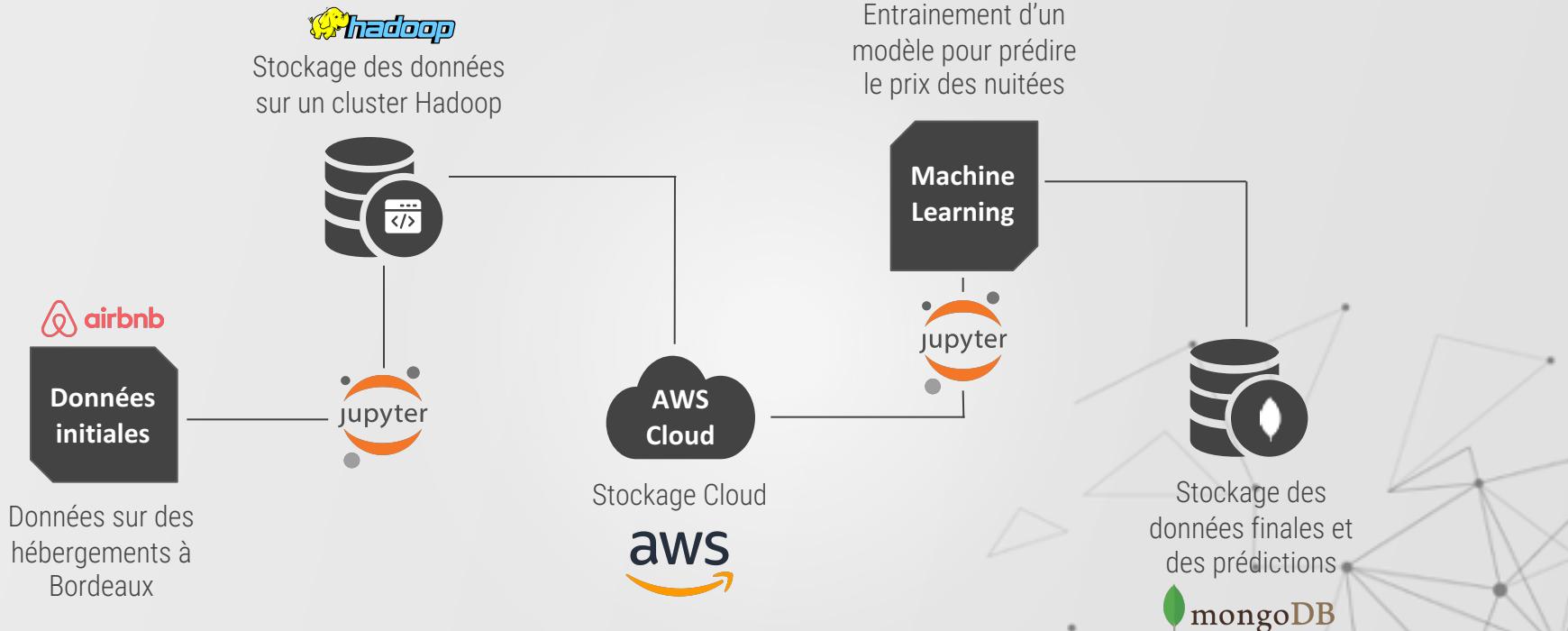
01

LE SUJET

Récapitulatif et précision des attentes



ANALYSE DU SUJET



02

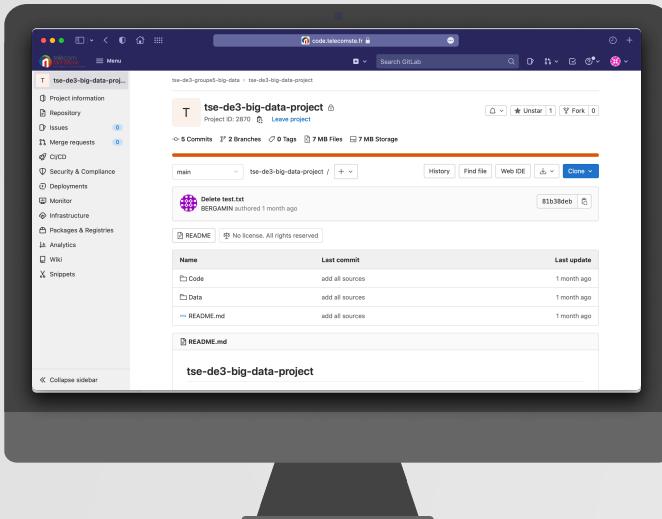
GESTION DE PROJET

Nos outils

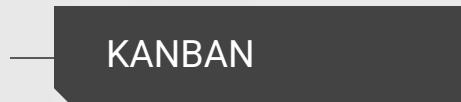


NOS OUTILS

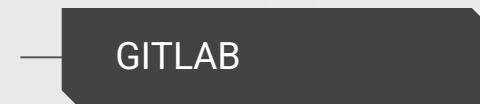
GANTT



KANBAN



GITLAB





03

NOTRE AVANCEMENT

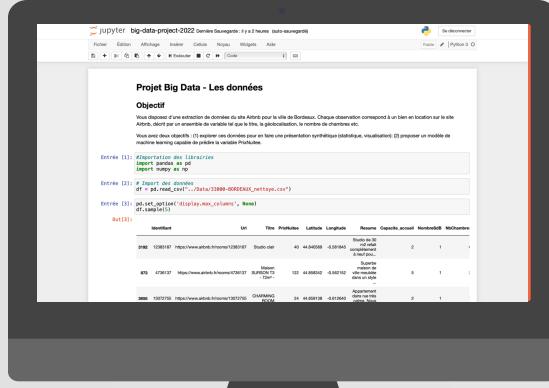
Tâches effectuées & problèmes rencontrés

ANALYSE ET TRAITEMENT DU JEU DE DONNÉES

Fichier .csv initial



Jupyter



The screenshot shows a Jupyter Notebook environment. The code cell contains:

```
# Import des bibliothèques
import pandas as pd
import numpy as np
```

The output cell shows the first few rows of a data frame named 'df':

Identifiant	URI	Préfixe	Latitude	Longitude	Reseau	Capacité accès	Nom de la cellule	
0	10901167	https://www.ubiqe.be/loc/10901167	Studio 010	40.4404598	-3.551645	0	1	0
1	47301167	https://www.ubiqe.be/loc/47301167	Studio 010	40.4404598	-3.551645	0	1	0
2	47301167	https://www.ubiqe.be/loc/47301167	Studio 010	40.4404598	-3.551645	0	1	0
3	10901167	https://www.ubiqe.be/loc/10901167	Studio 010	40.4404598	-3.551645	0	1	0



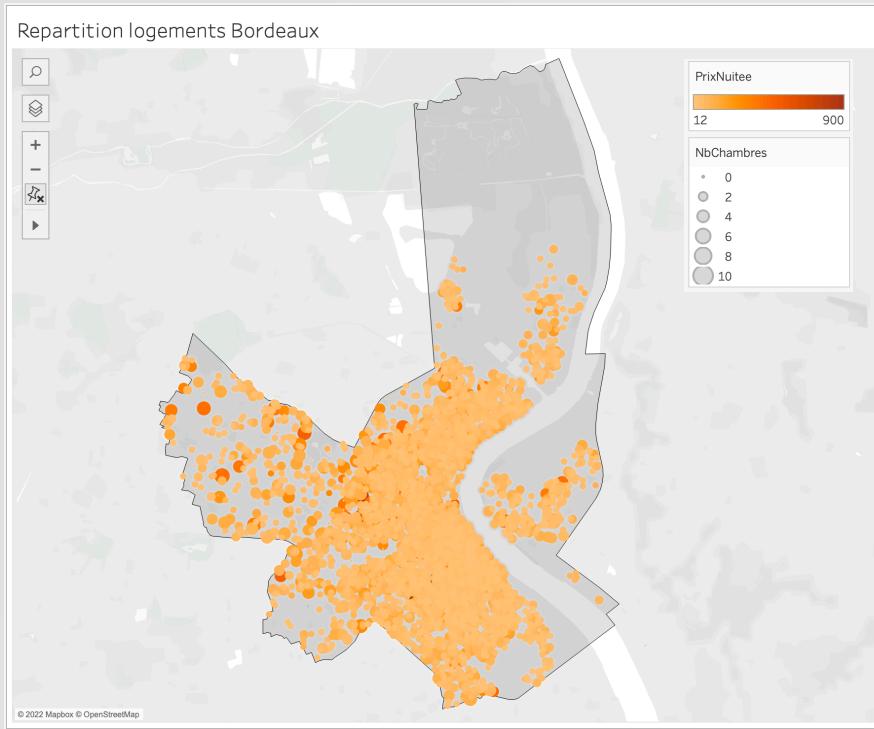
Post Traitement
Génération d'un nouveau
fichier de données

Renommage de colonnes

Suppression de données inutiles

Modification de valeurs abérantes

PREMIÈRE EXPLORATION DE DONNÉES



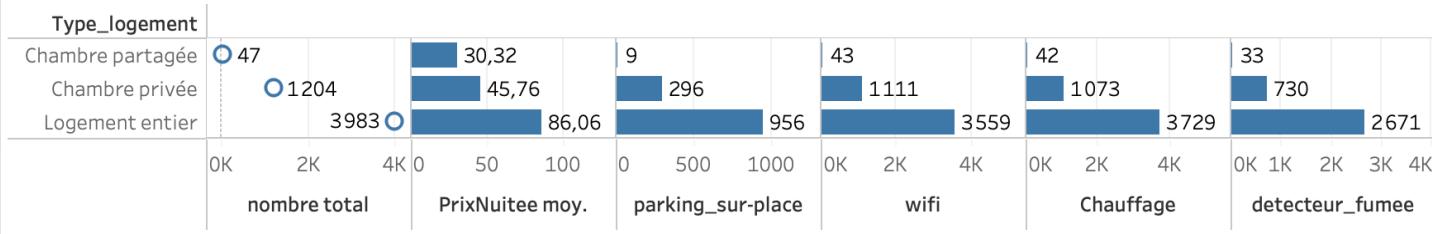
Pas de logements au nord de
Bordeaux

Forte concentration aux alentours
de la Gironde



PREMIÈRE EXPLORATION DE DONNÉES

Répartition par type de logement



Peu de chambres partagées

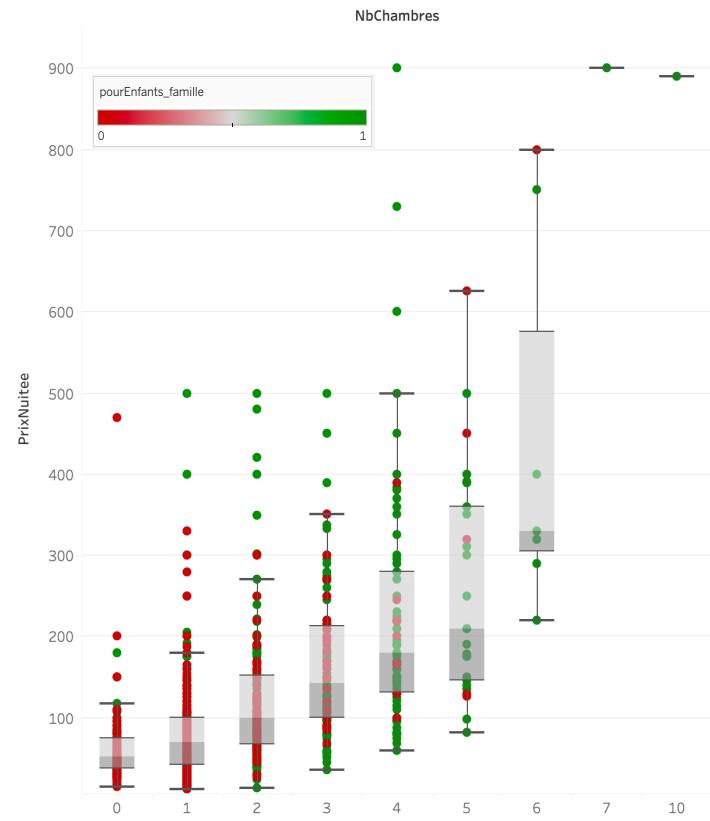
Peu de logements ont un parking sur place

Tous les logements n'ont pas de chauffage ou de détecteur de fumée



PREMIÈRE EXPLORATION DE DONNÉES

Repartition prix par nb chambre

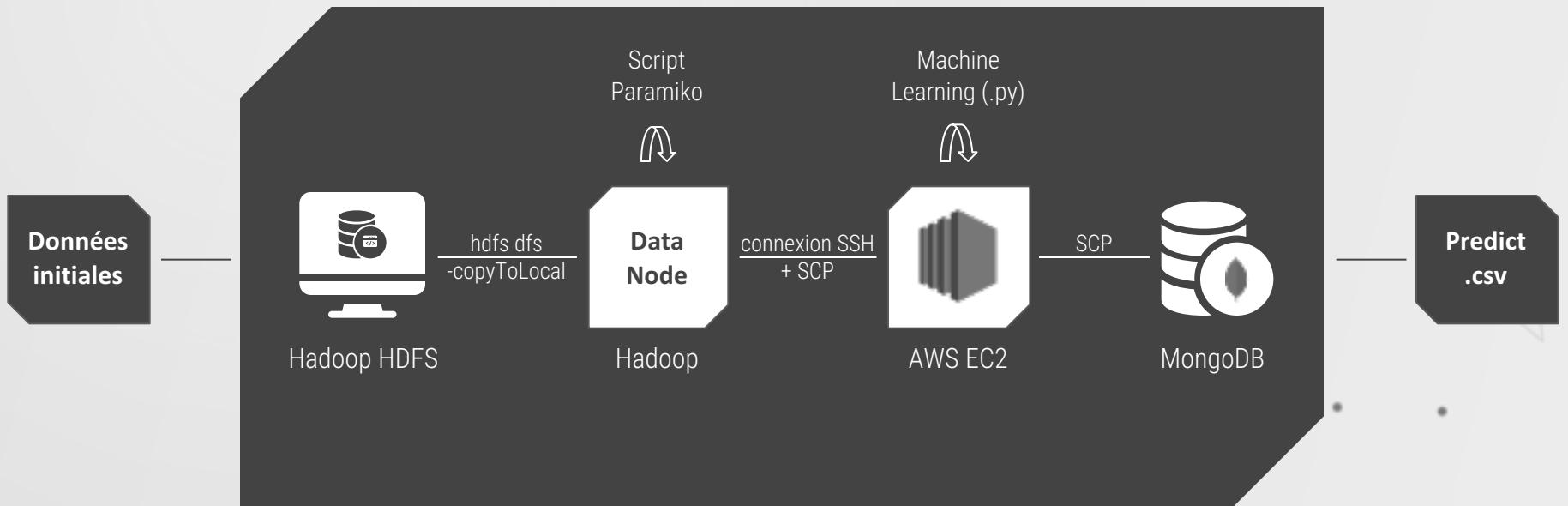


Peu de données sur les logements de 5 chambres et plus

Logements de 3 chambres et moins peu adaptés aux familles avec enfants



CONFIGURATION HADOOP & AWS



04

LA SUITE DU PROJET

Prochaines étapes

CONCLUSIONS DE L'EXPLORATION

VALEURS ABERRANTES

Supprimer les données ayant des prix $\geq 200\text{€}$

Nettoyage des données

dans le but d'augmenter l'efficacité du modèle d'apprentissage

TEXTE

Suppression des caractères spéciaux et chiffres pour ne garder que du texte

LANGAGE

Homogénéisation de la langue utilisée dans les données

CONFIGURATION HADOOP, HDFS & MONGODB



AUTOMATISATION

Automatisation des différents scripts pour industrialiser la solution

Création et développement du modèle d'apprentissage des données

AWS



MONGODB

Création de la base de données MySQL MongoDB



MERCI POUR VOTRE ATTENTION

BERGAMIN Maximilien – JAULGEY Thomas – REYNARD Thibaut
