

The background of the slide features a complex network graph on the left side, composed of numerous small dark grey dots connected by thin grey lines. To the right of the graph, several larger, semi-transparent white triangles of varying sizes are scattered across the white space.

BIG DATA PROJECT

Soutenance Finale

**BERGAMIN Maximilien
JAULGEY Thomas
REYNARD Thibaut**



RECAPITULATIF DU SUJET

01

GESTION DE PROJET

02

SOMMAIRE



03

REALISATIONS

04

DÉMONSTRATION



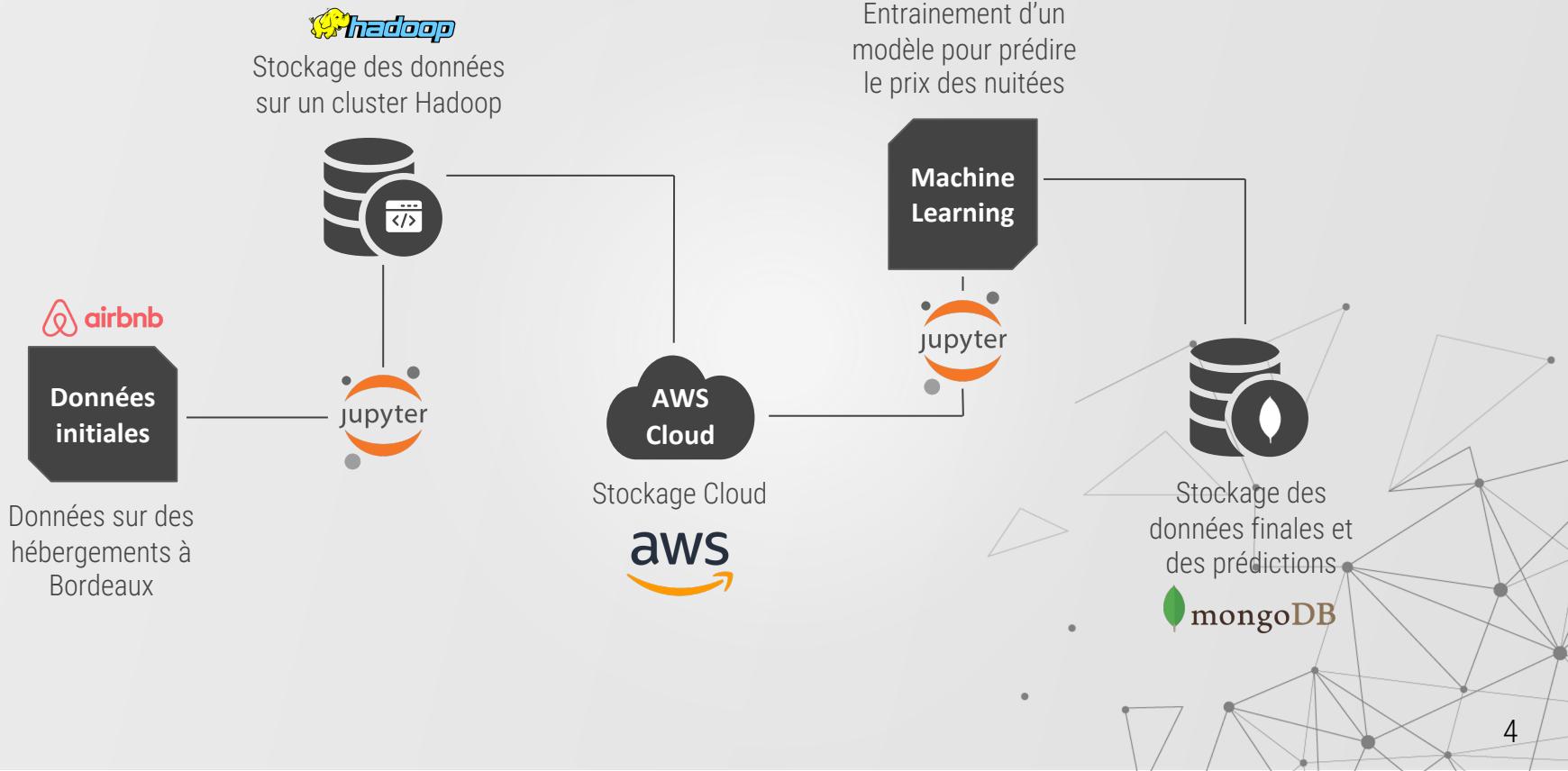
01

LE SUJET

Récapitulatif et précision des attentes



ANALYSE DU SUJET



02

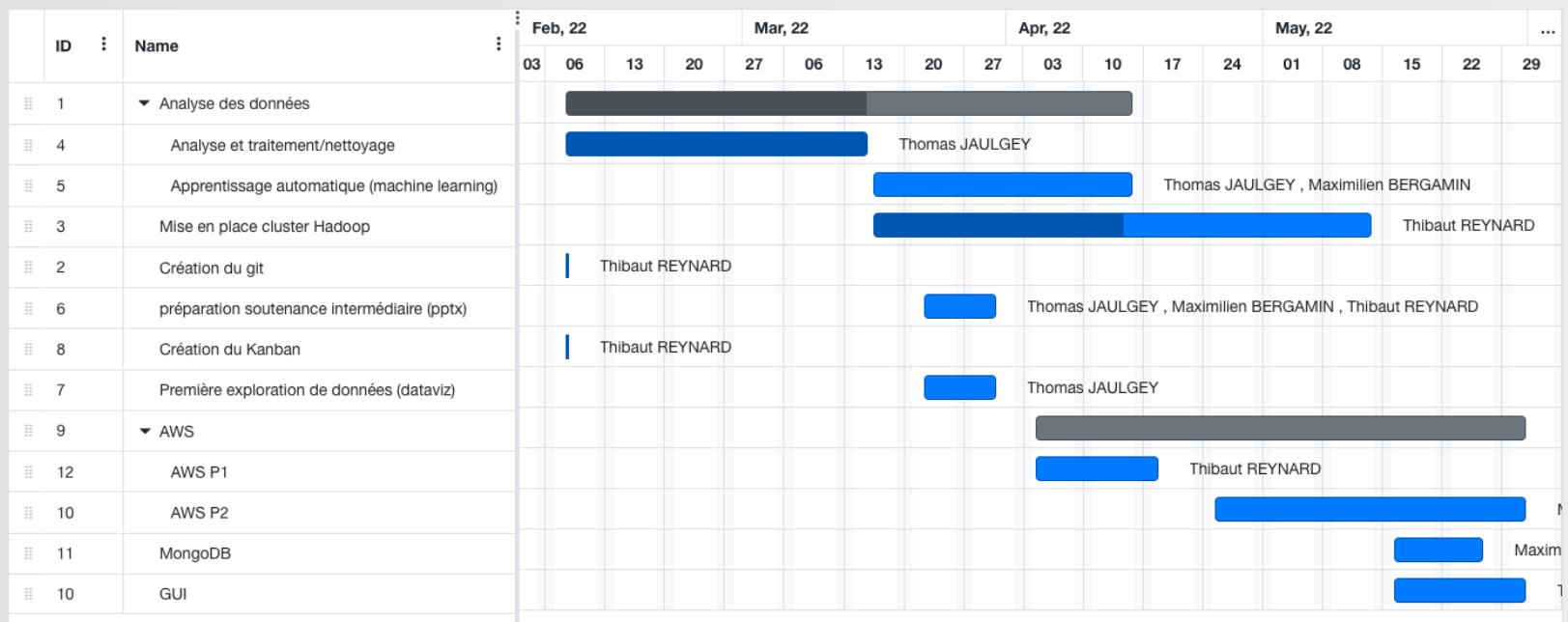
GESTION DE PROJET

Nos outils



NOS OUTILS

GANTT



NOS OUTILS

KANBAN

Tableau Kanban

Rechercher dans ce tab Version

Uniquement mes tickets Récemment mis à jour

BACKLOG 3	EN COURS 0	TERMINÉ 8
Automatisation de l'envoi des données prédites sur AWS = TDBDP-16		
Mise en place de la GUI dans le cloud = TDBDP-17		
Import automatique du fichier brut des données sur HDFS = TDBDP-18		
		Apprentissage automatique / machine learning = TDBDP-7
		Création GUI = TDBDP-15
		Création instance EC2 / Bucket S3 = TDBDP-13
		Mise en place du cluster Hadoop / Script de rapatriement des données = TDBDP-11
		Première exploration des données (Data viz sur Tableau) - TDBDP-9

NOS OUTILS

GITLAB

The screenshot shows a GitLab repository interface. The sidebar on the left lists various project sections: Project information, Repository (selected), Files, Commits, Branches, Tags, Contributors, Graph, Compare, Issues (0), Merge requests (0), CI/CD, Security & Compliance, Deployments, Monitor, Infrastructure, Packages & Registries, Analytics, Wiki, and Snippets. The main content area displays a merge request from 'REYNARD' to 'main' branch, authored 1 hour ago. Below it is a table of files with their last commit and update times:

Name	Last commit	Last update
Code	rename	1 hour ago
Data	add scripts and dataviz	1 hour ago
Soutenances/_intermediaire	update GUI et readme	5 days ago
Sujet	python termine et GUI ok	6 days ago
images	rename	1 hour ago
README.md	rename	1 hour ago
gantt.gantt	add scripts and dataviz	1 hour ago

Below the file list is a section titled 'tse-de3-big-data-project' containing the text: "© DE3 - BERGAMIN Maximilien - JAULGEY Thomas - REYNARD Thibaut". At the bottom, there are two sections: 'exécution de la GUI' with the instruction "se mettre dans le dossier /Code et exécuter python gui.py en ligne de commande" and 'affichage du gantt'.



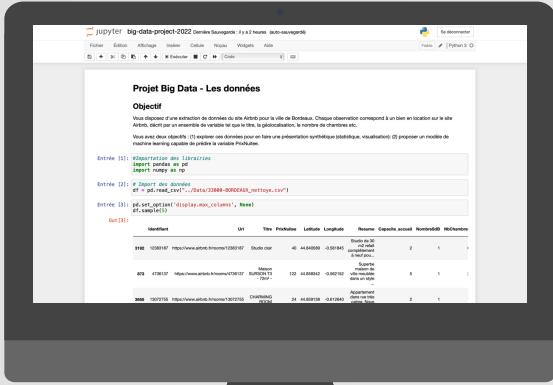
03

Réalisations

Tâches effectuées & problèmes rencontrés

ANALYSE ET TRAITEMENT DU JEU DE DONNÉES

Fichier .csv initial



Post Traitement
Génération d'un nouveau
fichier de données

Modification de valeurs abérantes

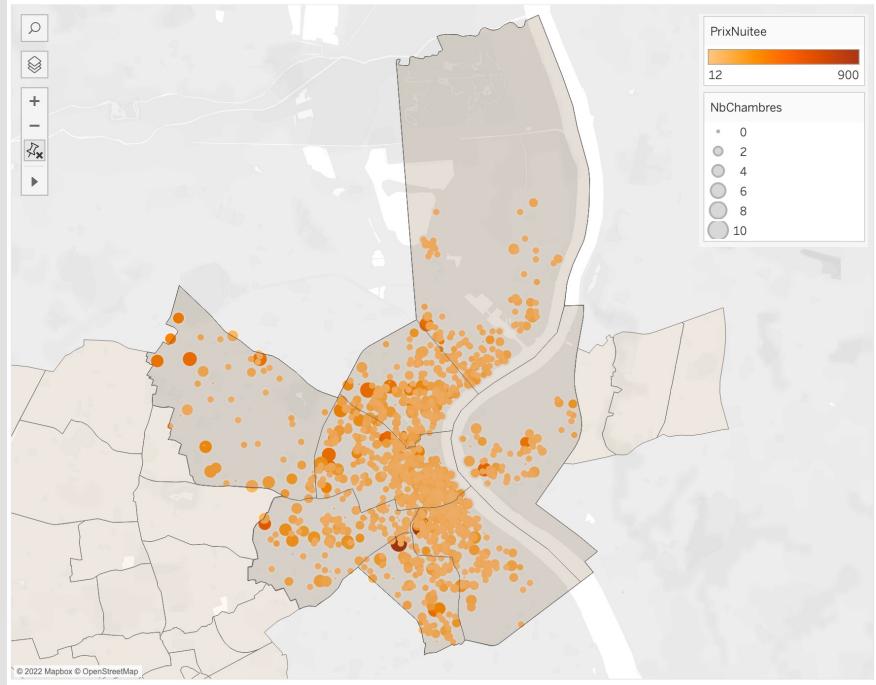
Renommage de colonnes / Suppression de données inutiles

Suppression des caractères spéciaux et chiffres

Homogénéisation de la langue utilisée dans les données

EXPLORATION DE DONNÉES

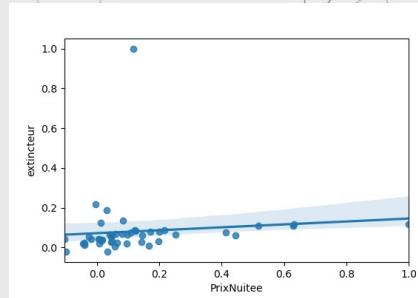
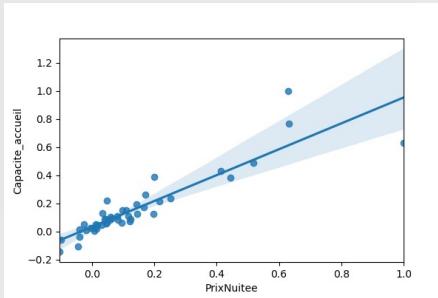
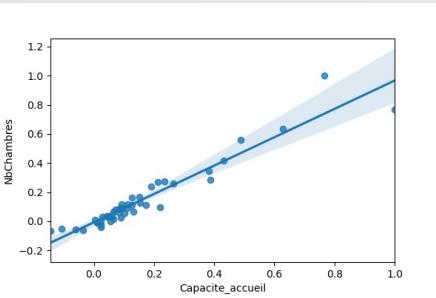
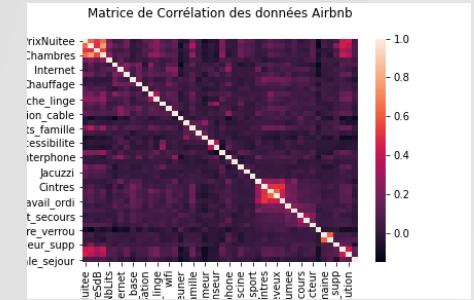
CAR - Répartition logements Bordeaux + filtres



Analyse géographique
Forte concentration aux alentours de la Gironde



EXPLORATION DE DONNÉES

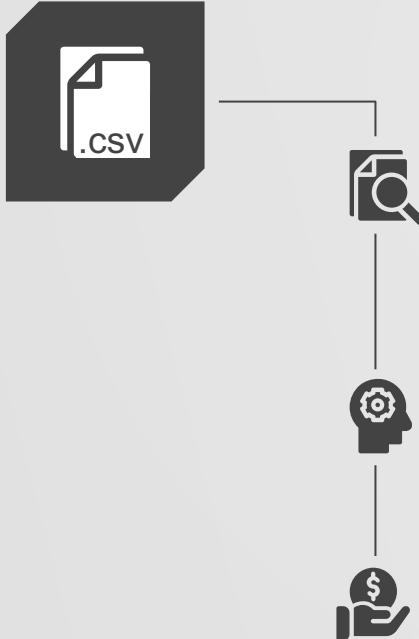


Etude des paramètres à utiliser dans le modèle

Matrice de corrélation
Regressions linéaires



ALGORITHME DE MACHINE LEARNING



Choix des colonnes

Quartier NombreSdB
NbLits Type_logement
Capacite_accueil Parking_sur_place

NbChambres
Type_propriete
Caution

Choix du modèle

xgboost regression lineaire erreur quadratique

Prédiction du prix d'une nuitée
erreur moyenne de 24,12€

airbnb - Prediction nuitée

Prédiction du prix d'une nuitée airbnb selon les critères définis

Quartier	Chartrons - Grand Parc - Jardin Public
Nb Salles de Bain	2
Nb Chambres	4
Nb Lits	6
Type Logement	Logement entier
Type Propriété	Villa
Capacité d'accueil	8
Parking sur place	Oui
Caution (€)	250

Générer un prix

Nous vous proposons de louer votre logement au prix de 141.24€ par nuitée.

GUI

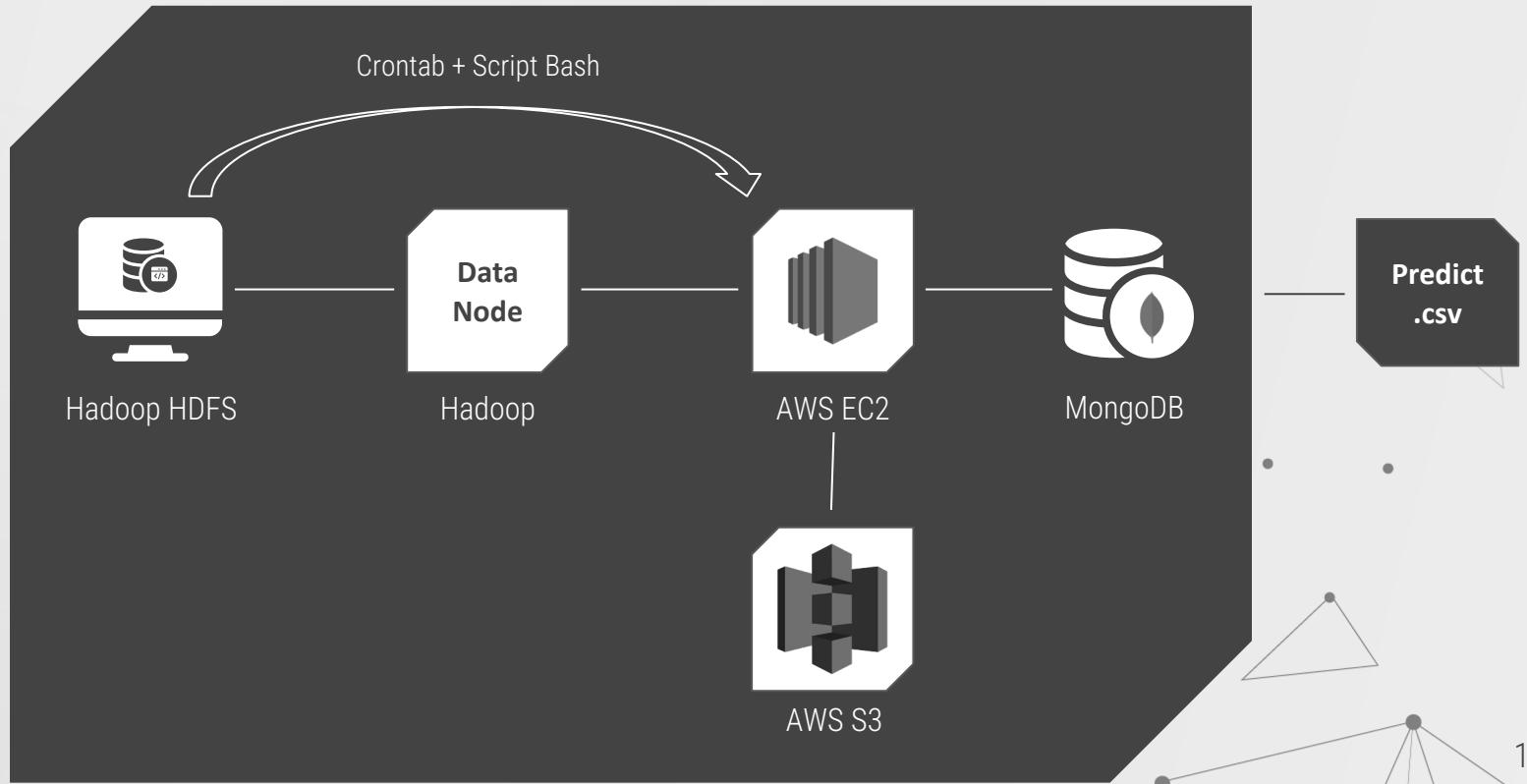
Interface graphique – Python – tkinter

Inputs correspondantes aux colonnes du modèle

Exécution de prediction.py au clic sur le bouton

Historisation sur mongodb des données prédictives

CONFIGURATION HADOOP & AWS



CONFIGURATION DE HADOOP

Name ➤	Size ➤	Last Modified ➤	Owner ➤	Group ➤
bigdata.csv	13.0 MB	2022-03-22 19:13	maria_dev	hdfs
key-big-data.pem	1.6 kB	2022-05-23 19:24	maria_dev	hdfs
script.sh	0.2 kB	2022-05-31 10:28	maria_dev	hdfs

Files View Ambari

Sandbox HDFS

```
[maria_dev@sandbox-hdp bigdata]$ ls  
key-big-data.pem  script.sh  
[maria_dev@sandbox-hdp bigdata]$
```

```
[maria_dev@sandbox-hdp bigdata]$ bash script.sh  
bigdata.csv  
[maria_dev@sandbox-hdp bigdata]$
```

```
#!/bin/bash --utf8  
sudo hdfs dfs -copyToLocal /tmp/data/bigdata.csv /bigdata  
sudo scp -i "key-big-data.pem" bigdata.csv ec2-user@ec2-54-160-240-133.compute-1.amazonaws.com:~/Data/bigdata.csv
```

Script bash

EC2 AWS

```
[ec2-user@ip-172-31-21-191 Data]$ ls  
bigdata.csv  predicted.csv  
[ec2-user@ip-172-31-21-191 Data]$
```

100% 13MB 738.0KB/s 00:18

```
55 23 * * * /usr/bin/sh /bigdata/script.sh  
~  
~  
~
```

Crontab

STOCKAGE CLOUD VIA AWS

```
//Création de l'instance EC2
var instanceParameters = {
  ImageId: 'ami-0022f774911c1d690',
  InstanceType: 't2.micro',
  KeyName: 'key-big-data',
  MinCount: 1,
  MaxCount: 1
}
```

```
D:\Telecom\Big-Data-2022>node scripts\script_aws.js bucket-big-data
```

Automatisation de la création d'une instance EC2 et d'un bucket S3

Permet de gagner du temps d'initialisation



STOCKAGE CLOUD VIA AWS

Détails | Sécurité | Mise en réseau | Stockage | Vérifications de statut | Surveillance | Balises

▼ Détails de l'instance Informations

Plateforme	ID AMI	Surveillance
Amazon Linux (dédou)	ami-0022f774911c1d690	désactivé
Informations sur la plateforme	Nom de l'AMI	Protection de la résiliation
Linux/UNIX	amzn2-ami-kernel-5.10-hvm-2.0.20220426.0-x86_64-gp2	Désactivé
Protection contre l'arrêt	Heure de lancement	Emplacement de l'AMI
Désactivé	Mon May 30 2022 11:41:17 GMT+0200 (heure d'été d'Europe centrale) (26 minutes)	amazon/amzn2-ami-kernel-5.10-hvm-2.0.20220426.0-x86_64-gp2
Récupération automatique de l'instance	Cycle de vie	Comportement Arrêt - Mise en veille prolongée
Par défaut	normal	désactivé
Index de lancement de l'AMI	Nom de la paire de clés	Motif de transition de l'état
0	key-big-data	-

bucket-big-data [Info](#)

Objets | Propriétés | Autorisations | Métriques | Gestion | Points d'accès

Objets (2)

Les objets sont les entités fondamentales stockées dans Amazon S3. Vous pouvez utiliser l'[inventaire Amazon S3](#) pour obtenir une liste de tous les objets de votre compartiment. Pour que d'autres personnes puissent accéder à vos objets, vous devez leur accorder explicitement des autorisations. [En savoir plus](#)

[Copier l'URI S3](#) [Copier l'URL](#) [Télécharger](#) [Ouvrir](#) [Supprimer](#) [Actions](#) [Créer un dossier](#) [Charger](#)

Rechercher des objets en fonction du préfixe

<input type="checkbox"/>	Nom	Type	Dernière modification	Taille	Classe de stockage
<input type="checkbox"/>	data/	Dossier	-	-	-
<input type="checkbox"/>	scripts/	Dossier	-	-	-

Instance et bucket créés
Configuration supplémentaire



STOCKAGE CLOUD VIA AWS

```
def mongoimport(csv_path, db_name, coll_name, db_url, db_port=27017):
    client = MongoClient(db_url, db_port)
    db = client[db_name]
    coll = db[coll_name]
    data = pd.read_csv(csv_path)
    payload = json.loads(data.to_json(orient='records'))
    coll.delete_many({})
    coll.insert_many(payload)
    print("Insert terminé")
```

```
import mongoimport

mongoimport.mongoimport("./Data/predicted.csv", "big-data-machine-learning", "predictions",
"mongodb+srv://mbergamin:ba6t32ms78tf@big-data-project.g26gj.mongodb.net/?retryWrites=true&w=majority")
```

```
import mongoimport

mongoimport.mongoimport("./Data/predicted.csv", "big-data-machine-learning", "donnees",
"mongodb+srv://mbergamin:ba6t32ms78tf@big-data-project.g26gj.mongodb.net/?retryWrites=true&w=majority")
```

```
0 0 * * * sudo wget https://bucket-big-data.s3.amazonaws.com/data/predicted.csv && mv predicted.csv Data/predicted.csv
0 0 * * * aws s3 mv Data/bigdata.csv s3://bucket-big-data/data/bigdata.csv
5 0 * * * python3 script/predicted.py
5 0 1,15 * * * python3 script/donnees.py
```



Scripts python pour envoyer prédictions sur MongoDB

Crontab pour exécuter ce script tous automatiquement



BASE DE DONNEES MONGODB

big-data-machine-learning

DATABASE SIZE: 18.44MB INDEX SIZE: 216KB TOTAL COLLECTIONS: 2

CREATE COLLECTION

Collection Name	Documents	Documents Size	Documents Avg	Indexes	Index Size	Index Avg
donnees	5237	18.46MB	3.61KB	1	180KB	180KB
predictions	2	514B	257B	1	36KB	36KB

QUERY RESULTS: 1-2 OF 2

```
_id:ObjectId("628e3b66a15f3cdc050ce7c0")
Quartier:"Centre ville"
NombreSdb:1
NbChambres:2
NbLits:2
tokenized_Type_logement:"logement entier"
tokenized_type_propriete:"appartement"
Capacite_accueil:4
parking_sur-place:1
Caution:140
prediction:64.4040908813
```

```
_id:ObjectId("628e3b66a15f3cdc050ce7c1")
Quartier:"Centre ville"
NombreSdb:3
NbChambres:4
NbLits:4
tokenized_Type_logement:"logement entier"
tokenized_type_propriete:"appartement"
Capacite_accueil:4
parking_sur-place:1
```

Une base de données
Deux collections



The background features a complex network of thin gray lines forming numerous triangles and dots of varying sizes, creating a sense of depth and connectivity.

04

DÉMONSTRATION



MERCI POUR VOTRE ATTENTION

BERGAMIN Maximilien – JAULGEY Thomas – REYNARD Thibaut