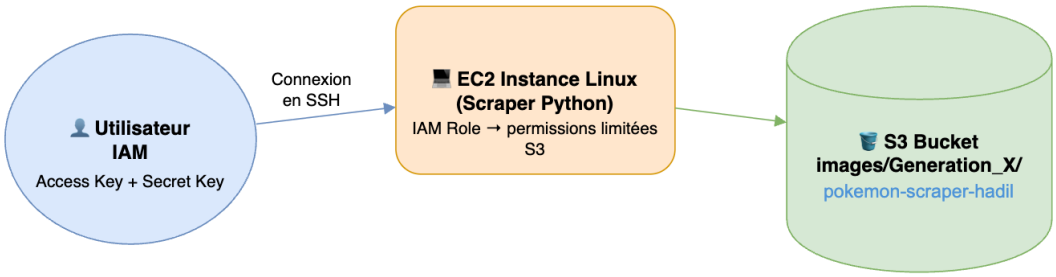
Rapport TP

1. Architecture mise en place

- **Utilisateur** IAM (Access Key + Secret Key) : déclenche l'exécution du script Python.
- **Instance EC2 (Amazon Linux)** : exécute le scraper écrit en Python
- **Bucket Amazon S3** : reçoit les images uploadées depuis l'EC2. Les fichiers sont organisés dans le préfixe images/<Generation>/

Le schéma d'architecture a été réalisé avec draw.io (lien dans GIT):



2. Choix techniques

- **Source des données** : Bulbapedia (lien) : la liste complète des Pokémons.
- **Scraping** : utilisation de requests (pour télécharger le HTML) et BeautifulSoup (pour extraire les URLs d'images). Les images sont organisées par génération (I, II, III...).
- **Stockage** : envoi direct des images dans S3 via le SDK boto3. Les fichiers suivent la structure: s3://pokemon-scraper-hadil/images/Generation_I/001_Bulbasaur.png
- **Historique des données :** Pour chaque image, le script fait une requête vers S3 pour vérifier si le fichier existe. S'il existe, il l'ignore sinon le récupère.
- **Gestion des erreurs** : try/except pour gérer les erreurs réseau (timeouts, échecs de connexion).

  Gestion des erreurs côté S3 (permissions, bucket introuvable).

- **Respect des bonnes pratiques** :

  time.sleep(1) entre chaque requête pour respecter le robots.txt.

  User-Agent explicite pour identifier le scraper.

  Pas de clés AWS en dur dans le code → utilisation d'IAM + aws configure.

  IAM avec les permissions minimales (s3:PutObject, s3:GetObject, s3:ListBucket)

3. Vidéo Démo IAM → EC2 → S3 : lien vers la video

4. Exemple Lien publique :

https://pokemon-scraper-hadil.s3.eu-north-1.amazonaws.com/images/Generation_VIII/0851_Centiskorch.png

https://pokemon-scraper-hadil.s3.eu-north-1.amazonaws.com/images/Generation_V/0496_Servine.png

5. Captures d’écran

- Politique pour l’utilisateur IAM

- Connexion SSH

```
(base) hadilneji@Mac Downloads % ssh -i "pokemon-key.pem" ec2-user@13.61.105.17


   ,     #_
   ~\_  ####_        Amazon Linux 2023
  ~~  \_#####\
  ~~     \###|
  ~~       \#/ ___   https://aws.amazon.com/linux/amazon-linux-2023
   ~~       V~' '->
    ~~~         /
      ~~._.   _/
         _/ _/
       _/m/'
[ec2-user@ip-172-31-40-248 ~]$
```

- Configuration AWS

```
[ec2-user@ip-172-31-40-248 ~]$ aws configure
AWS Access Key ID [****************s/qY]: AKIAUACVYTSTRRX6RWAY
AWS Secret Access Key [****************s/qY]: fhzJT7D7VRERwx7pAW5WG4f3FvxU2jsxrn
2Fs/qY
Default region name [eu-north-1b]: eu-north-1
Default output format [json]: json
[ec2-user@ip-172-31-40-248 ~]$ aws s3 ls
```

- Exécution du script python pour la 1ère fois

```
[ec2-user@ip-172-31-40-248 ~]$ nano scraper-pokemon1.py
[ec2-user@ip-172-31-40-248 ~]$ python3 scraper-pokemon1.py
2025-09-04 08:58:05,574 - INFO - Scraping Generation_I ...
2025-09-04 08:58:05,675 - INFO - Uploadé sur S3 → s3://pokemon-scraper-hadil/images/Generation_I/0001_Bulbasaur.png
2025-09-04 08:58:06,729 - INFO - Uploadé sur S3 → s3://pokemon-scraper-hadil/images/Generation_I/0002_Ivysaur.png
2025-09-04 08:58:07,794 - INFO - Uploadé sur S3 → s3://pokemon-scraper-hadil/images/Generation_I/0003_Venusaur.png
2025-09-04 08:58:08,859 - INFO - Uploadé sur S3 → s3://pokemon-scraper-hadil/images/Generation_I/0004_Charmander.png
2025-09-04 08:58:09,919 - INFO - Uploadé sur S3 → s3://pokemon-scraper-hadil/images/Generation_I/0005_Charmeleon.png
2025-09-04 08:58:10,981 - INFO - Uploadé sur S3 → s3://pokemon-scraper-hadil/images/Generation_I/0006_Charizard.png
2025-09-04 08:58:12,043 - INFO - Uploadé sur S3 → s3://pokemon-scraper-hadil/images/Generation_I/0007_Squirtle.png
2025-09-04 08:58:13,098 - INFO - Uploadé sur S3 → s3://pokemon-scraper-hadil/images/Generation_I/0008_Wartortle.png
2025-09-04 08:58:14,155 - INFO - Uploadé sur S3 → s3://pokemon-scraper-hadil/images/Generation_I/0009_Blastoise.png
2025-09-04 08:58:15,218 - INFO - Uploadé sur S3 → s3://pokemon-scraper-hadil/images/Generation_I/0010_Caterpie.png
```

- Exécution du script python pour la 2ème fois (si l’image existe -> SKIP)

```
2025-09-04 12:05:07,391 - INFO - Scraping Generation_I ...
2025-09-04 12:05:07,444 - INFO - [SKIP] 0001_Bulbasaur.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:08,460 - INFO - [SKIP] 0002_Ivysaur.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:09,477 - INFO - [SKIP] 0003_Venusaur.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:10,494 - INFO - [SKIP] 0004_Charmander.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:11,510 - INFO - [SKIP] 0005_Charmeleon.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:12,528 - INFO - [SKIP] 0006_Charizard.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:13,545 - INFO - [SKIP] 0007_Squirtle.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:14,561 - INFO - [SKIP] 0008_Wartortle.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:15,576 - INFO - [SKIP] 0009_Blastoise.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:16,591 - INFO - [SKIP] 0010_Caterpie.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:17,608 - INFO - [SKIP] 0011_Metapod.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:18,624 - INFO - [SKIP] 0012_Butterfree.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:19,648 - INFO - [SKIP] 0013_Weedle.png existe déjà dans S3, pas d'écrasement.
2025-09-04 12:05:20,666 - INFO - [SKIP] 0014_Kakuna.png existe déjà dans S3, pas d'écrasement.
```

- Données dans **Bucket S3**

```
[ec2-user@ip-172-31-40-248 ~]$ aws s3 ls s3://pokemon-scraper-hadil/  --recursiv
e
2025-09-04 09:54:57      10239 images/Generation_I/0001_Bulbasaur.png
2025-09-04 09:54:58      10446 images/Generation_I/0002_Ivysaur.png
2025-09-04 08:58:08       9099 images/Generation_I/0003_Venusaur.png
2025-09-04 09:00:46       7673 images/Generation_I/0151_Mew.png
2025-09-04 09:00:47       6386 images/Generation_II/0152_Chikorita.png
2025-09-04 09:00:48       6595 images/Generation_II/0153_Bayleef.png
2025-09-04 09:00:49       6795 images/Generation_II/0154_Meganium.png
2025-09-04 09:00:50       7069 images/Generation_II/0155_Cyndaquil.png
```

- Accès publics seulement au dossier image