DOSSIER D'ARCHITECTURE TECHNIQUE

Version 1.0

30/02/2022

Historique

Création du document : 30/02/2022

Groupe 1

GONCALVES Swan

CHEN Michel

NIEV Jimmy

2022 – 2023

Sommaire

Table des matières

Présentation générale

1 • Présentation du client

Inetum est une entreprise de services et de solutions digitales (ESN). Présent dans plus de 27 pays, le Groupe compte près de 28 000 collaborateurs, elle opère dans plusieurs secteurs d'activité, fournissant des solutions spécifiques sur mesure pour répondre aux besoins spécifiques pour chaque industrie :

- Banque & Finance
- Santé
- Industrie
- Télécommunications et médias
- Distribution
- Energie

2 • Présentation de l'équipe

Nous sommes une start-up spécialisée dans les solutions "Machine Learning" : TriniTech. Composée de 3 Data ingénieurs, qui sont :

- Swan Gonçalves
- Michel Chen
- Jimmy Niev

Nous proposons la conception et la réalisation d'environnement de projet "Machine Learning".

3 • Contexte de développement

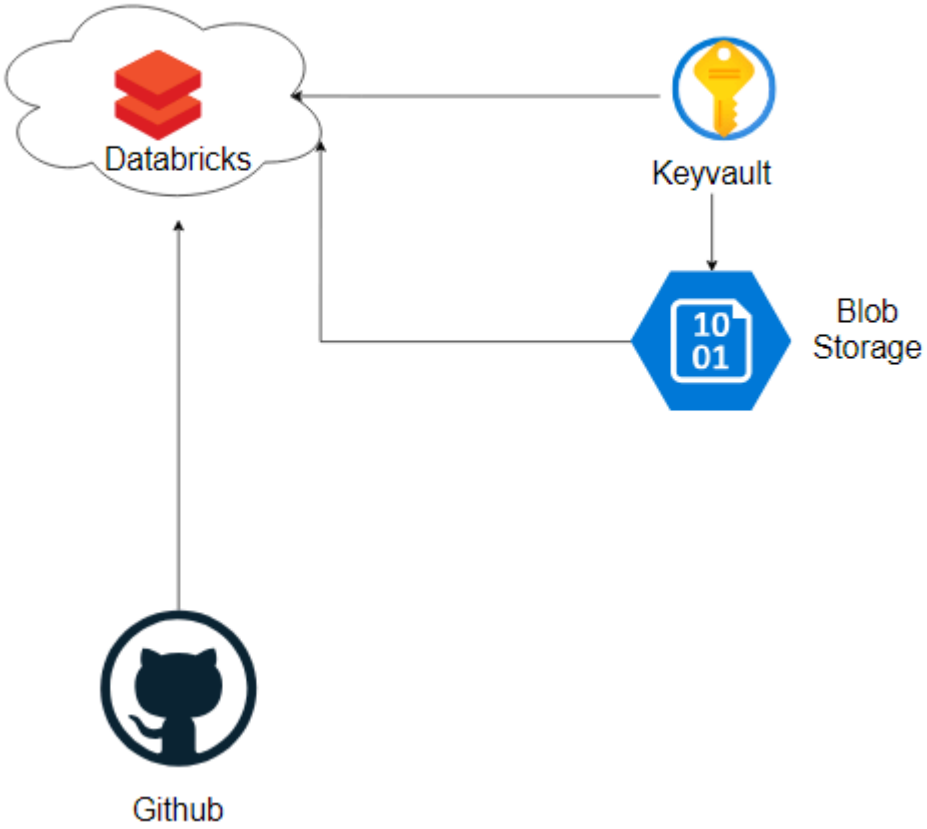
Nous avons été dépêchés par notre client - Inetum - afin de réaliser un modèle de machine learning permettant de prédire la catégorie énergétique d'un logement, ceci afin de faciliter le travail des différents agents du secteur de l'immobilier.

Architecture applicative

1 • Liste des composants applicatifs

- **Databricks :** Plateforme d'analyse de données et d'IA, utilisée pour le traitement et l'analyse des données.
- **Keyvault :** Gestion des secrets et des clés, utilisée pour stocker et gérer de manière sécurisée les secrets, tels que les clés API, les mots de passe, etc.
- **Blob storage :** Stockage des données, utilisé pour stocker un grand volume de données non structurées, telles que du texte ou des images binaires.
- **Github** : Gestion du code source et CI/CD, utilisé pour le stockage, la gestion et le suivi du code source, ainsi que potentiellement pour orchestrer les pipelines CI/CD.

2 • Schéma détaillé de l'architecture applicative



Architecture Technique

1 • Liste ressources (exemple)

Dataset: DPE

- **Source** : Hackaton ESGI x Inetum - Kaggle
- **Taille** : 718 Mo
- **Format** : csv
- **Besoin de retravailler** : La donnée nous est fournie en brute.
    - Une première séparation en train, test, validation a été faite, mais nécessite de
    - Colonnes en français, parfois illisible
    - Vectorisation de nos colonnes
    - Beaucoup de colonnes en doublons,
    - Fusion de colonnes, Data Cleaning
- **Enrichissement** : La donnée est assez complète pour notre besoin, mais au besoin nous pouvons la compléter grâce aux portail open data de l'ADAME

Dans cette même ressource est fournie le **data-dictionary** détaillant les colonnes de notre dataset.

2 • Listes détaillées des blobs

3 • Liste des bases de données

- Base de données d'entraînement
- Base de données de validation

4 • Listes des tables

Nous aurons 2 tables :

- Notre table DPE raw qui contiendra l'ensemble de nos données en brute
- Notre table DPE cleaned qui contiendra l'ensemble de nos données une fois nettoyées.

5 • UML

DPE
Id
DPE_Number
Heating_Installation_Configuration_No2
Entered_Solar_Coverage_Factor
Living_Area_Served_by_DHW_Installation
Lighting_GHG_Emissions
Stairwell
Final_Energy_Consumption_5_Uses_No2
Cold_Generator_Type
Emitter_Type_Heating_Installation_No2
Total_Surface_Photovoltaic_Sensors
Municipality_Name_Raw
Wasteful_Heating_Consumption_Heating_Installation_No1
Heating_Cost_Energy_No2
GHG_Emissions_Heating_Energy_No2
INSEE_Code_Raw
Energy_Type_No3
GHG_Label
Generator_Type_No1_Installation_No2
Postal_Code_Raw
Description_Generator_Heating_No2_Installation_No2
Solar_Coverage_Factor
Construction_Year
Altitude_Class
Postal_Code_Clean
Final_Consumption_5_Uses_per_m2
Final_Consumption_5_Uses
DPE_Label
Ceiling_Height
Department_Number_Clean
Envelope_Insulation_Quality
Joinery_Insulation_Quality
Wall_Insulation_Quality
Lower_Floor_Insulation_Quality
Upper_Floor_Insulation_Quality_Converted_Attic
Upper_Floor_Insulation_Quality_Unconverted_Attic
Upper_Floor_Insulation_Quality_Roof_Terrace
Building_Living_Area
Building_Type