MISE EN PLACE D'UN DATA LAKE POUR L'ANALYSE DU MARCHÉ DU TRAVAIL

BOUAZIZI M., CASANOVA S. , CONDE K. , RAMOS Y.

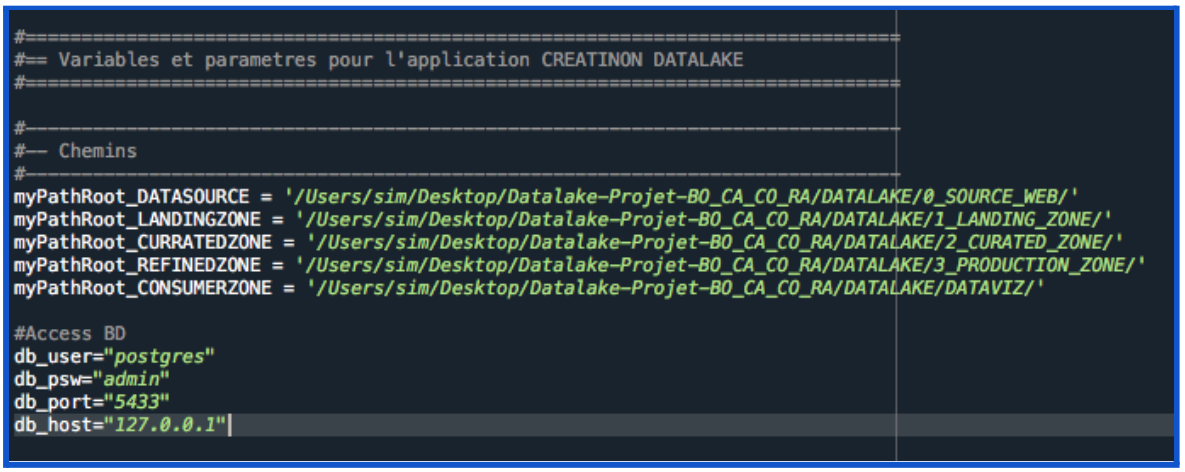
Afin de pouvoir exécuter le projet ***"Mise en place d'un data lake de recherche d'emploi"***, il faut suivre les instructions suivantes, si vous rencontrez quelques difficultés, vous pourriez écrire à savacano_28@live.com.

Récupération de projet

1. Pour récupérer les sources, il faut aller sur : **https://drive.google.com/drive/folders/1FHWRjaHH5Zpck65COK6CBcIv_-5d7-jH?usp=sharing**
   Dans ce repository vous trouverez :
   - Rapport en pdf -> **Datalake-Rapport.pdf**
   - Lisez-moi en pdf -> **Lisez-moi.pdf**
   - Dataviz - PowerBI en pbix et pdf -> **Dataviz-PowerBI.pbix**
   - Dump postgres en dump -> **datalake-postgres.dump**
   - Projet python estructuré selon les etapes : source_web, landing_zone, curated_zone, production_zone, dataviz, dvlp -> **Dossier Datalake-Projet-BO_CA_CO_RA**

Exécution du projet

1. Pour exécuter le projet il faudrait initialement configurer les suivant informations dans le fichier **DATALAKE/DVLP/PYTHON/Datalake_Parametrage.py:**
   -Chemin de dossier
   - Accès à la base de données : login, password et port
   Exemple :



```
#==========================================================================
#== Variables et parametres pour l'application CREATINON DATALAKE
#==========================================================================

#--------------------------------------------------------------------------
#-- Chemins
#--------------------------------------------------------------------------
myPathRoot_DATASOURCE = '/Users/sim/Desktop/Datalake-Projet-BO_CA_CO_RA/DATALAKE/0_SOURCE_WEB/'
myPathRoot_LANDINGZONE = '/Users/sim/Desktop/Datalake-Projet-BO_CA_CO_RA/DATALAKE/1_LANDING_ZONE/'
myPathRoot_CURRATEDZONE = '/Users/sim/Desktop/Datalake-Projet-BO_CA_CO_RA/DATALAKE/2_CURATED_ZONE/'
myPathRoot_REFINEDZONE = '/Users/sim/Desktop/Datalake-Projet-BO_CA_CO_RA/DATALAKE/3_PRODUCTION_ZONE/'
myPathRoot_CONSUMERZONE = '/Users/sim/Desktop/Datalake-Projet-BO_CA_CO_RA/DATALAKE/DATAVIZ/'

#Access BD
db_user="postgres"
db_psw="admin"
db_port="5433"
db_host="127.0.0.1"
```

2. Après, il faudrait modifier le fichier **DATALAKE/DVLP/BATCH/Lancer_Mise_En_Place_datalake.bat** afin de donner les chemins correspondant à **Python** et le fichier d'exécution du projet **Datalake_Programme_Principal.py.**

Exemple :

Lancer_Mise_En_Place_Datalake.bat

```
1  echo     MISE EN PLACE DATALAKE DONNESS MARCHE DE TRAVAIL
2  echo         M2 BI-BD - Universite Lyon2
3  echo         Cours G.D.M. - TD DATALAKE
4  echo               2020-2021
5  echo    DEBUT MISE EN PLACE DATALAKE
6  python  /Users/sim/Desktop/Datalake-Projet-BO_CA_CO_RA/DVLP/PYTHON/Datalake_Programme_Principal.py
7  echo    FIN
```

3. Ensuite, il faudrait donner de permis d'exécution (ou tous) avec le commande **chmod 777 Lancer_Mise_En_Place_datalake.bat**
   Exemple :

```
(base) ➜  BATCH ls
Lancer_Mise_En_Place_Datalake.bat
(base) ➜  BATCH ./Lancer_Mise_En_Place_Datalake.bat
zsh: permission denied: ./Lancer_Mise_En_Place_Datalake.bat
(base) ➜  BATCH sudo chmod 777 Lancer_Mise_En_Place_Datalake.bat
Password:
(base) ➜  BATCH ./Lancer_Mise_En_Place_Datalake.bat
```

4. Finalement, vous pourrez lancer le projet avec **./Lancer_Mise_En_Place_Datalake.bat**
   Exemple :

```
(base) ➜  BATCH ./Lancer_Mise_En_Place_Datalake.bat
MISE EN PLACE DATALAKE DONNESS MARCHE DE TRAVAIL
M2 BI-BD - Universite Lyon2
Cours G.D.M. - TD DATALAKE
2020-2021
DEBUT MISE EN PLACE DATALAKE
**** Debut du traitement en Secondes 1616939814.1508899 ***
Landing Zone
/Users/sim/Desktop/Datalake-Projet-BO_CA_CO_RA/DATALAKE/0_SOURCE_WEB/
/Users/sim/Desktop/Datalake-Projet-BO_CA_CO_RA/DATALAKE/0_SOURCE_WEB/
/Users/sim/Desktop/Datalake-Projet-BO_CA_CO_RA/DATALAKE/0_SOURCE_WEB/
Curated Zone
Rafinnage Zone
Table SOCIETE created successfully
Table EMPLOI created successfully
Table FAIT_EMPLOIS created successfully
Table AVIS created successfully
Table FAIT_AVIS created successfully
Les insertions dans la table societe ont été faits
[nltk_data] Downloading package stopwords to /Users/sim/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
Les insertions dans la table emploi ont été faits
Les insertions dans la table avis ont été faits
Les insertions dans la table Fait_Avis ont été faits
Les insertions dans la table Fait_Emplois ont été faits


**** Fin du traitement en Secondes 1616939929.3914728 ***

**** Duree du traitement en Secondes 115.24058294296265 ***

FIN
(base) ➜  BATCH
```

Mise en place de la base de données sur postgres

1. Nous avons travaillé avec postgres pour sauvegarder les données et les résultats des analyses, c'est pour cela qu'il est nécessaire d'installer un postgres version > 12.0.
   Gardez les identifiants de : user, password et port.
2. Après de l'installation, il faut télécharger et restaurer la base de données **"BASE_CURATED_ZONE"** dans votre postgres.
   Le fichier à utiliser est : **"Datalake-postgres.dmp"**, et la commande à utiliser est : **pg_restore "BASE_CURATED_ZONE" < *Datalake-postgres.dmp***.
   Si vous avez une interface graphique, il faudrait juste créer la bd **"BASE_CURATED_ZONE"**, et chercher l'option : restore en utilisant le fichier dump.

Visualisation de dashboard en PowerBI

1. Il faudrait installer **PowerBi dektop** ou se connecter à **powerbi.microsoft.com** et charger le fichier **Dataviz-PowerBI.pbix.**