Rapport de Projet Data - Qualité de l'eau (LTM)

Frik Elias

Rafael Nakache

Ossama Loridi

Zerguit Manel

1. Présentation du Projet

Objectif

L'objectif de ce projet est de concevoir une architecture de type **Lakehouse** pour intégrer, transformer, modéliser et analyser des données issues du programme **Long Term Monitoring (LTM)**, portant sur la qualité de l'eau dans divers états américains.

Données sources

Nous avons utilisé trois fichiers CSV fournis par un organisme environnemental :

- Site_Information_2022_8_1.csv : informations sur les sites de prélèvement (coordonnées, état, etc.).
- Methods_2022_8_1.csv : méthodologies scientifiques d'analyse des paramètres.
- LTM_Data_2022_8_1.csv : mesures relevées sur le terrain.

2. Architecture du Projet & Modèle Conceptuel

Nous avons mis en place une architecture **Lakehouse** organisée en trois couches :

Bronze (staging)

- Lecture brute des CSV avec gestion du schéma et encodage.

Silver (curated layer)

- Nettoyage, transformation, uniformisation des types.
- Implémentation des **Slowly Changing Dimensions (SCD)** :
    - **SCD Type 1** pour les dimensions stables (ex : site).
    - **SCD Type 2** pour les dimensions à historique (ex : méthode).

Gold (analytique)

- Données finalisées en **format long**, prêtes pour Power BI.
- Table de faits avec enrichissement par jointures dimensionnelles.

3. Traitements effectués (PySpark)

a. Nettoyage des données (Silver)

```
valeurs_aberrantes = ["", "NA", "null", "-1"]

for col in df.columns:

   df = df.withColumn(col, F.when(F.col(col).isin(valeurs_aberrantes), None).otherwise(F.col(col)))
```

- Conversion : float, int, timestamp
- Uniformisation des noms de paramètres via regexp_replace

b. Transformation en format long

```
param_cols = [...] # liste des colonnes métriques


# Pivot

df_long = df.selectExpr("SITE_ID", "PROGRAM_ID", ..., f"stack({len(param_cols)}, ...)")
```

c. Gestion des SCD

```
window = Window.partitionBy("PROGRAM_ID", "PARAMETER").orderBy(F.desc("END_YEAR"))

df_method_scd1 = df_method.withColumn("row_num", F.row_number().over(window)).filter("row_num = 1")
```

d. Jointures

```
df_joint = df_fait_long.join(df_method, on=["PROGRAM_ID", "PARAMETER"], how="left")

df_joint = df_joint.join(df_site.drop("PROGRAM_ID"), on="SITE_ID", how="left")
```

e. Enregistrement dans la table finale

```
df_gold.write.format("delta").mode("overwrite").saveAsTable("gold_water_quality")
```

4. Tables et Types de SCD

Table	Type	SCD	Description
silver_dim_site	Dimension	1	Localisation et info site
silver_dim_method_scd1	Dimension	1	Dernière méthode connue pour un paramètre
silver_dim_method_scd2	Dimension	2	Historique des méthodes

silver_ltm_data_scd1	Faits	1	Mesure la plus récente
silver_ltm_data_scd2	Faits	2	Historique complet
gold_water_quality	Faits	-	Jointures sur les dimensions

5. Visualisations Power BI

1. Carte des mesures par site géographique

- **Objectif** : visualiser la répartition géographique des mesures.
- **Données** : LATDD, LONDD, COUNT(VALUE)

2. Histogramme - Nombre de mesures par paramètre

- **Objectif** : identifier les paramètres les plus fréquemment mesurés.
- **Données** : PARAMETER, COUNT(VALUE)

3. Évolution annuelle des mesures

- **Objectif** : suivre l'évolution de la surveillance par année.
- **Données** : year, COUNT(VALUE)

6. Référentiel de Code

Tous les scripts de transformation sont disponibles sur : **GitHub** :

7. Conclusion

Dans ce projet vous aller retrouver toutes ces etapes :

- Intégration de **3 sources distinctes**.
- Architecture **Lakehouse** robuste.
- Modèle en étoile facilitant l'analyse.
- Gestion de **2 types de SCD** (1 & 2).
- Visualisations pertinentes et adaptées aux objectifs de suivi environnemental.

Rendu final :

- Rapport PDF (ce document)
- Table Delta finale gold_water_quality
- Rapport Power BI
- Repository Git contenant les scripts de traitement