Evaluation et amélioration de la qualité de donnée de consommation énergétique

Projet

M2 DataScale

Zoubida Kedad



1

Utilisation d'un ETL pour évaluer et améliorer la qualité des données

- Mise en œuvre d'un ETL
 - Hop (https://hop.apache.org/)
 - Airflow (https://airflow.apache.org/)
 - NiFi (https://nifi.apache.org/)
 - Pentaho (https://pentaho.com/pentaho-developer-edition/)
 - Talend
- Définition d'un scénario comportant
 - Des processus d'évaluation de la qualité des données
 - Des processus d'amélioration de la qualité des données

Zoubida Kedad

2

2

Scénario : intégration de données de consommation énergétique

- Plusieurs sources de données distinctes
- Des objets cibles qui seront calculés à partir des sources de données

Zoubida Kedad

3

3

Le schéma cible – Consommation d'énergie

Consommation_IRIS_Paris (<u>ID_IRIS</u>, Conso_moyenne_annuelle)

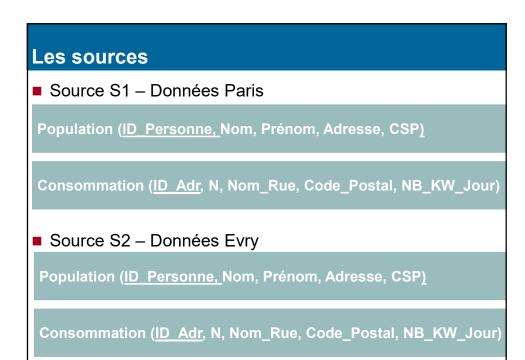
Consommation_IRIS_Evry (<u>ID_IRIS</u>, Conso_moyenne_annuelle)

Consommation_CSP (<u>ID_CSP</u>, Conso_moyenne_annuelle, Salaire_Moyen)

Zoubida Kedad

4

Δ



Zoubida Kedad

5

Les sources (suite) Source S3 – Catégories socio-professionnelles CSP (ID_CSP, Desc, Salaire_Moyen, Salaire_Min, Salaire_Max) Source S4 – IRIS (Ilots Regroupés pour l'Information Statistique) IRIS (ID_Rue, ID_Ville, ID_Iris)

6

Problèmes de qualité

- Conformité à un format, une codification
- Hétérogénéité des échelles, de la granularité
- Complétude des données
- Détection et élimination de doublons

Zoubida Kedad

7

7

Résultats attendus

- Bilan sur l'audit de la qualité des sources et des données calculées
 - Pour les facteurs de qualité considérés
- Workflow de traitement permettant d'améliorer la qualité des données en utilisant l'ETL choisi
- Date de fin : semaine du 17/11/2025
- Livrables
 - Compte rendu oral par groupe pendant les séances de cours
 - Démonstration
 - Compte rendu écrit

Zoubida Kedad

3

8