Qualité des données

Etude de cas

M2 DataScale

Zoubida Kedad

UNIVERSITÉ DE VERSAILLES SAINT-QUENTIN-EN-YVELINES

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

- Mise en œuvre d'un ETL
  - Talend
- Définition d'un scénario comportant
  - Des processus d'évaluation de la qualité des données
  - Des processus d'amélioration de la qualité des données

Scénario : intégration de données de pollution

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

Le schéma cible – Pollution de l'air

Paris (ID_Polluant, Taux_moyen_jour, NBS, NBC, Statut)

Hauts_de_Seine (ID_Polluant, Taux_moyen_jour, NBS, NBC, Statut)

Yvelines (ID_Polluant, Taux_moyen_jour, NBS, NBC, Statut)

Stations (ID_Station, Adresse, Tel, Contact_mail)

- NBS : nombre de stations utilisées pour le calcul
- NBC : nombre de capteurs mobiles utilisés pour le calcul
- Statut : « Alerte_pollution » ou « Normal »

Les sources

- Source S1 – Mesures de stations fixes

Mesures (ID_Polluant, Date, ID_Station, Taux_relevé)

Station (ID_Station, Num, Rue, Ville, Code_postal, Tel, Contact_Mail)

- Source S2 – Mesures de stations fixes

Mesures (ID_Polluant, Date, ID_Station, Taux_relevé)

Station (ID_Station, Num, Rue, Ville, Code_postal, Tel, Contact_Mail)

Les sources (suite)

- Source S3 – Polluants

Polluants (ID_Polluant, Description, Seuil_toléré)

- Source S4 – Mesures de capteurs mobiles

Mesures (ID_Polluant, Date, ID_Capteur, Localisation, Taux_relevé)

- Source S5 – Mesures de capteurs mobiles

Mesures (ID_Polluant, Date, ID_Capteur, Localisation, Taux_relevé)

Problèmes de qualité

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

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 : 12/11/2021
- Livrables
  - Compte rendu oral par groupe pendant les séances de cours
  - Démonstration
  - Compte rendu écrit (**pas de rapport !**)