

PROJET QUALITÉ DE DONNÉES

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

Réalisé par :
OUNIS Wafa
ZARROUKI Wafa
EL MIR Omar
BOUZID Achref

Encadré par : Mme Zoubida KEDAD

2019/2020

Scénario:

► Sources:

- – Fichier XLS 
- Fichier CSV 
- Fichier XML 



► Cible :

- –Entrepôt de données



université
PARIS-SACLAY

ÉTUDIANT

(IDÉtudiant, Nom, Prénom, Nationalité, téléphone, Email, Adresse, Salaire)

- **ÉTUDIANT** (IDÉtudiant, Nom, Prénom, Nationalité, NumTelephone, Email, Adresse, Revenus)

Hétérogénéité des échelles, de la granularité

- S_i : \min_i , \max_i , mean_i , Ratio_i
- $\text{MeanG} = \frac{\sum \text{Mean}_i}{3}$
- $\text{Ratio}_i = \frac{\text{mean}_i}{\text{MeanG}}$
- $\text{Seuil} = 1/\sqrt{1000}$
- Si $\text{Ratio}_i < \text{seuil}$  problème de granularité au niveau de la source i
 $\text{Revenus} = \text{Revenus} * 1000$

Source 1: Université de Versailles

numEtu	Nom	Prenom	Salaire
21806957	BARRY	oumar salamata	2.11
21805387	CAMARA	ousmane	0.827



IDEtudiant	Nom	Prenom	Revenus
21806957	BARRY	oumar salamata	2110
21805387	CAMARA	ousmane	827

Les Doublons

Source 1: Université de Versailles

numEtu	Nom	Prenom	Nationalite	Tel	Email	Salaire	CodePostal
21565256	Ounis	Wafa	Tunisie	09 28 68 77 21	wafa.ounis@ens.uvsq.fr	1	95000

Source 2: Université de Paris Sud

NumEtudiant	Nom	Prenom	Nationalite	NumTelephone	Adresse	Email	Revenus
600489624	zarrouki	wafa	Tunisie	08 44 04 86 77	78140		
600798568	El-Mir	Omar	Tunisie	07 14 31 10 58	95250		22

Source 3: Université d'Evry

```
<etudiant>
  <etuID>813244376</etuID>
  <Nom>Ounis</Nom>
  <Prenom>Wafa</Prenom>
  <Nationalite>Tunisie</Nationalite>
  <Telephone>+33 5 46 10 54 04</Telephone>
  <Adresse>95000</Adresse>
  <Email>Wafa.Ounis@univ-evry.fr</Email>
  <Salaire>1</Salaire>
</etudiant>
```

```
<etudiant>
  <etuID>815655994</etuID>
  <Nom>El-Mir</Nom>
  <Prenom>Omar</Prenom>
  <Nationalite>Tunisie</Nationalite>
  <Telephone>+33 7 14 31 10 58</Telephone>
  <Adresse>95250</Adresse>
  <Email> </Email>
  <Salaire>22.00</Salaire></etudiant>
```

```
<etudiant>
  <etuID>811123467</etuID>
  <Nom>zarrouki</Nom>
  <Prenom>wafa</Prenom>
  <Nationalite>Tunisie</Nationalite>
  <Telephone>+33 5 68 29 52 06</Telep
  <Adresse>78140</Adresse>
  <Email> </Email>
  <Salaire>184</Salaire>
```



Amélioration:

IDEtudiant	Nom	Prenom	Nationalite	Telephone	Adresse	Email	Revenus
21565256,813244376	Ounis	Wafa	Tunisie	09 28 68 77 21	95000	wafa.ounis@ens.uvsq.fr	1
600798568,815655994	El-Mir	Omar	Tunisie	07 14 31 10 58	95250		22
600489624,811123467	zarrouki	wafa	Tunisie	08 44 04 86 77	78140		184

Complétude des données

Source 2: Université de Paris Sud

NumEtudiant	Nom	Prenom	Revenus
600508568	Quintessa	Orli	
	Dustin	Jasper	
	Abbot	Allistair	

« IDEtudiant »
ou « Revenus »
n'ont pas de
valeurs

Amélioration:

Supprimer
les tuples qui
n'ont pas
d'IDEtudiant

Revenus
= 0

IDEtudiant	Nom	Prenom	Revenus
600508568	Quintessa	Orli	0

Conformité à un format

Source 1: Université de Versailles

numEtu	Nom	Prenom	Tel	Email
21908045	ILYASS	HAMZA	04 3	h@fr
21004407	DJOUDI	fares	09 73	fares@fr

Email = « nom.prenom@Domaine_Univ.fr »
Telephone = NULL

Amélioration:

IDEtudiant	Nom	Prenom	Telephone	Email
21908045	ILYASS	HAMZA		HAMZA.ILYASS@ens.uvsq.fr
21004407	DJOUDI	fares		fares.DJOUDI@ens.uvsq.fr