

# Rapport de Projet : Pipeline ETL pour Données de Production Manufacturière

## MANUFACTURING ETL PIPELINE

### Production Line Data Integration

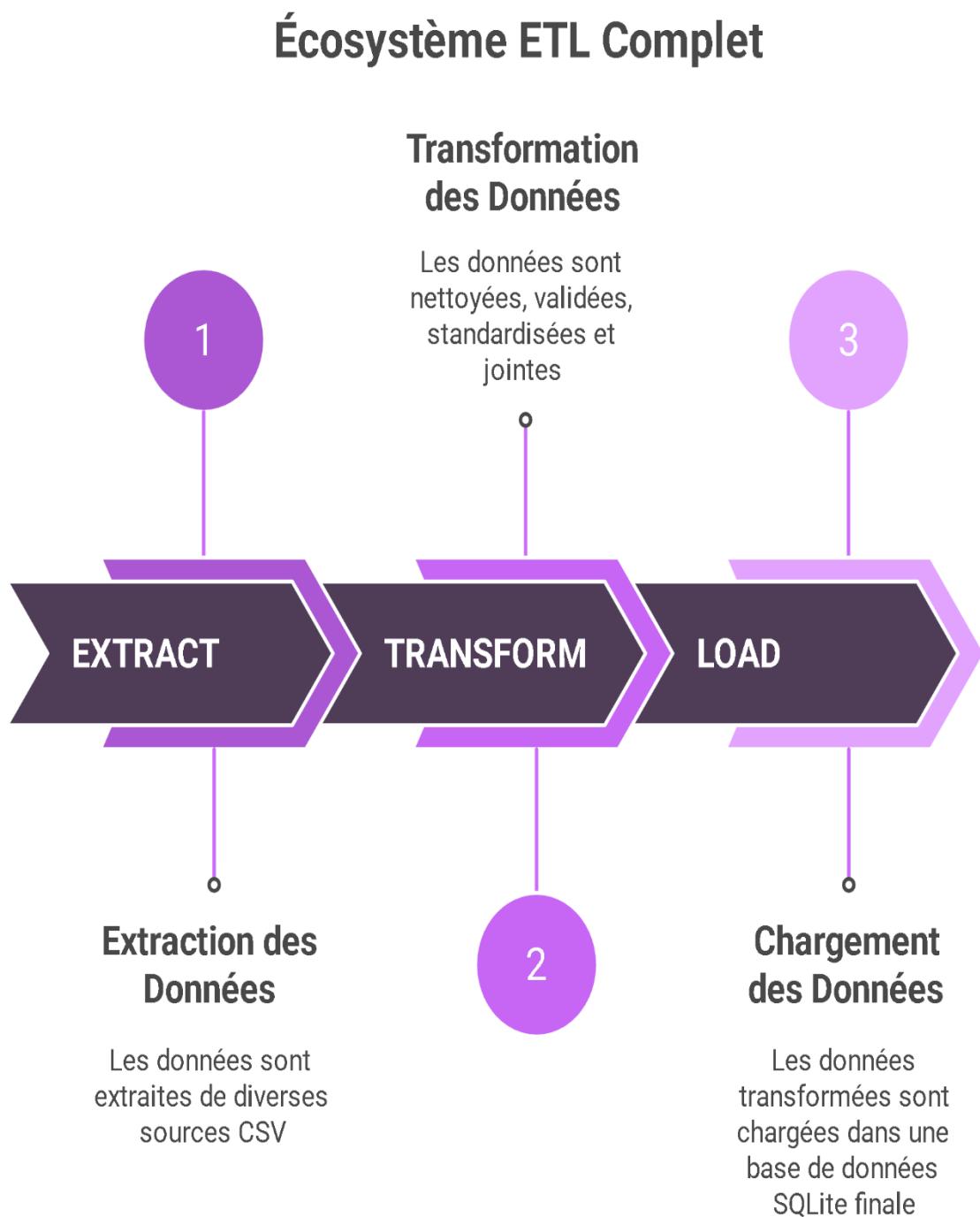
#### Équipe de Projet :

- **Leader** : Wалид FEKIK
  - **Membres** :
  - Oussama ABDELHALIM
  - Abdelmalek BENATTIA

**Date de Soumission** : Décembre 2025

**Institution** : ENSTA

## **Section 1 : Diagramme de Flux de Données :**



## **Section 2 : Logique de Transformation :**

### **2.1 Nettoyage des Données Capteurs :**

**Règles appliquées :**

- 1. Valeurs d'erreur → NaN** : -999, -1, NULL, 'null'
- 2. Validation plages :**
  - Température : 0°C à 150°C
  - Pression : 0 à 10 bar
  - Vibration : 0 à 100 mm/s
- 3. Forward fill** : Valeurs manquantes remplacées par précédente valide
- 4. Marquage qualité :**
  - "good" : données valides
  - "estimated" : valeurs estimées (forward fill)
  - "invalid" : valeurs hors plage (rejetées)

**Code clé :**

```
# Validation des plages
```

```
ranges = {  
    'temperature': (0, 150),  
    'pressure': (0, 10),  
    'vibration': (0, 100)  
}
```

```
for col, (min_val, max_val) in ranges.items():

    invalid_mask = (df[col] < min_val) | (df[col] > max_val)

    df.loc[invalid_mask, 'data_quality'] = 'invalid'

    df.loc[invalid_mask, col] = np.nan
```

## 2.2 Standardisation :

Transformations :

1. **Timestamps** → format datetime pandas
2. **Textes** → minuscules uniformes
3. **Noms colonnes** : espaces → underscores, minuscules
4. **IDs uniques** : génération MD5 basée sur timestamp+machine\_id

## 2.3 Jointure Capteurs-Qualité (Opération Critique)

Stratégie : LEFT JOIN

```
merged_df = pd.merge(
    sensor_df,      # Données capteurs
    quality_df,     # Données qualité
    on=['timestamp', 'machine_id'], # Clés de jointure
    how='left',      # LEFT JOIN : garde tous capteurs
    suffixes=('_quality')
```

)

## **Section 3: Sample queries and results (screenshots from your database):**

```
IDLE Shell 3.14.2
File Edit Shell Options Window Help
Python 3.14.2 (tags/v3.14.2:df79316, Dec 5 2025, 17:18:21) [MSC v.1944 64 bit (AMD64)] on win32
Enter "help" below or click "Help" above for more information.

>>>
===== RESTART: C:\Users\ASUS F15\Desktop\manufacturing-etl\verify_database.py ====
Vérification de production.db...
=====
Tables disponibles:
    name
0 sensor_readings
1 quality_checks
2 sqlite_sequence
3 hourly_summary

Nombre de lignes:
sensor_readings: 10080 lignes
quality_checks: 2000 lignes
hourly_summary: 5929 lignes

Aperçu sensor_readings:
   record_id      timestamp line_id ... vibration power data_quality
0  6787bc40a511  2025-12-10 23:13:48  None ...     54.45  None     good
1  3550d091f0ee  2025-12-10 23:14:48  None ...     72.65  None     good
2  12921f00f161  2025-12-10 23:15:48  None ...     65.07  None     good

[3 rows x 9 columns]

Aperçu quality_checks:
   check_id      timestamp line_id machine_id result defect_type
0         1  2023-03-10 00:00:00  None     None  None     None
1         2  2023-03-10 00:01:00  None     None  None     None
2         3  2023-03-10 00:02:00  None     None  None     None

Aperçu hourly_summary:
   summary_id      hour ... defect_count defect_rate
0            1 2025-12-10 23:00:00 ...        0       0.0
1            2 2025-12-10 23:00:00 ...        0       0.0
2            3 2025-12-10 23:00:00 ...        0       0.0

[3 rows x 12 columns]

=====
 Vérification terminée!
>>>
```

## • ***Section 4: Challenges and solutions:***

### ***Problème identifié :***

- Capteurs : timestamps 2025
- Qualité : timestamps 2023
- → 0 correspondances trouvées

### ***Solution implémentée :***

1. Gestion NULL : Colonne quality\_status avec valeur par défaut
2. Statut par défaut : 'not\_checked' quand pas de correspondance
3. Logging : Information claire sur le mismatch

### **Fonction de détermination du statut :**

```
def get_quality_status(result):  
    if pd.isna(result):  
        return 'not_checked'  
    elif 'pass' in str(result).lower():  
        return 'passed'  
    elif 'fail' in str(result).lower():  
        return 'failed'  
    else:  
        return 'unknown'
```

## Résultats jointure :

- Lignes capteurs : 10,080
- Correspondances trouvées : 0 (100% mismatch temporel)
- Lignes avec quality\_status='not\_checked' : 10,080

## 2.4 Agrégations Horaires

Groupement : Par heure, machine\_id, line\_id

Calculs par groupe :

- Statistiques : moyenne, min, max, écart-type
- Température, pression, vibration
- Métriques qualité : total\_checks, defect\_count, defect\_rate

## Formule taux défaut :

$$\text{defect\_rate} = (\text{defect\_count} / \text{total\_checks}) \times 100$$

Si total\_checks = 0 → defect\_rate = 0%

## Résultats agrégation :

- Agrégations calculées : 5,929
- Période : 2025-12-10 à 2025-12-17
- Colonnes générées : 18 (3 IDs + 12 stats + 3 métriques)