

I3T

Préparation des données SCADA pour le projet de prévision ML

Réalisé par : Yesmine Srairi

Encadré par : Sahbi Gsouma

28 juillet 2025

Introduction

Dans le cadre de mon stage chez **I3T**, j'ai été chargée de préparer les données issues du système SCADA afin de permettre leur exploitation dans un projet de prévision du niveau d'eau dans un réservoir à l'aide de l'intelligence artificielle (IA).

Les données brutes présentaient de nombreuses lacunes : valeurs manquantes, absence de cohérence temporelle, mesures dispersées sur plusieurs tables, noms de colonnes ambigus, ou encore horodatages imprécis ou désordonnés. Pour garantir la fiabilité des analyses, une étape préalable de traitement BI s'est imposée : nettoyage, restructuration, uniformisation des noms, et agrégation temporelle.

Ce travail a permis de produire des vues consolidées, exploitables pour la modélisation prédictive et l'analyse intelligente des niveaux d'eau.

Nettoyage des données : Cas de la station de Kairouan

Étant donné les contraintes de temps et la priorité donnée à une mise en œuvre rapide de la phase de prévision, le nettoyage des données a été effectué exclusivement sur les données issues de la station de Kairouan.

Ce choix a permis de concentrer les efforts sur une station représentative du réseau, tout en garantissant un niveau de qualité suffisant pour amorcer les travaux d'analyse prédictive.

Le processus de nettoyage a consisté à :

- Identifier et supprimer les valeurs manquantes ;
- Uniformiser les noms de colonnes pour une meilleure lisibilité ;
- Filtrer les données selon leur station d'origine ;
- Filtrer les variables inutilisables ou non informatives.

À l'issue de ce traitement, les données de Kairouan ont pu être utilisées pour construire des vues horaires exploitables par les outils de modélisation.

Les requêtes ont été exécutées principalement sur les vues `View_Information`, `View_ArchivedInformation` et `View_Stations`.

1. Analyse de la densité et de la régularité des mesures : exemple du réservoir Ain Bidha

L'exemple suivant illustre l'analyse effectuée pour la station associée au réservoir Ain Bidha (`STA_SiteNumber` = 3), permettant de mesurer la densité des données, la présence de valeurs nulles, et la période couverte. **Des analyses similaires ont été menées pour les trois vues principales.**

Listing 1 – Analyse de la densité des mesures pour le réservoir Ain Bidha

```
SELECT
    vi.INF_Label ,
    vi.INF_Unit ,
    COUNT(vai.INF_Value) as record_count ,
    MIN(vai.INF_Date) as earliest_date ,
    MAX(vai.INF_Date) as latest_date ,
    COUNT(DISTINCT CAST(vai.INF_Date AS DATE)) as days_with_data ,
    SUM(CASE WHEN vai.INF_Value IS NULL THEN 1 ELSE 0 END) as
        null_count ,
    ROUND(COUNT(vai.INF_Value) * 100.0 /
        NULLIF(DATEDIFF(hour , MIN(vai.INF_Date) , MAX(vai.INF_Date)
        )), 0), 2) as data_density_percent
FROM dbo.View_Informationes vi
LEFT JOIN dbo.View_ArchivedInformations vai
    ON vi.STA_SiteNumber = vai.STA_SiteNumber
    AND vi.INF_NumberInStation = vai.INF_NumberInStation
WHERE vi.STA_SiteNumber = 3
GROUP BY vi.INF_Label , vi.INF_Unit
ORDER BY record_count DESC;
```

	INF_Label	INF_Unit	record_count	earliest_date	latest_date	days_with_data	null_count	data_density_percent
1	Débit Dist DN250	L/s	1045817	2021-06-10 10:04:00.00000000	2025-07-09 10:40:00.00000000	1472	0	2924.5400000000000
2	Débit Dist DN100	L/s	1045815	2021-06-10 10:04:00.00000000	2025-07-09 10:40:00.00000000	1472	0	2924.5400000000000
3	NIVEAU	m	1045812	2021-06-10 10:06:00.00000000	2025-07-09 10:40:00.00000000	1472	0	2924.5300000000000
4	Signal gsm		975444	2021-09-20 09:12:00.00000000	2025-07-09 10:40:00.00000000	1372	0	2928.1200000000000
5	CTR DIST-DN100	m3	34637	2021-06-10 11:00:00.00000000	2025-07-09 10:00:00.00000000	1467	0	96.8600000000000
6	DN100-Productio(24h)	m3	34637	2021-06-10 11:00:00.00000000	2025-07-09 10:00:00.00000000	1467	0	96.8600000000000
7	DN100-Productio(h)	m3	34637	2021-06-10 11:00:00.00000000	2025-07-09 10:00:00.00000000	1467	0	96.8600000000000
8	DN250-Productio(24h)	m3	34637	2021-06-10 11:00:00.00000000	2025-07-09 10:00:00.00000000	1467	0	96.8600000000000
9	DN250-Productio(h)	m3	34637	2021-06-10 11:00:00.00000000	2025-07-09 10:00:00.00000000	1467	0	96.8600000000000
10	CTR DIST-DN250	m3	34636	2021-06-10 11:00:00.00000000	2025-07-09 10:00:00.00000000	1467	0	96.8600000000000
11	Défaut Equipement		11775	2021-06-10 17:13:27.00000000	2025-07-09 04:59:43.00000000	1267	0	32.9400000000000
12	Niveau Très Haut		2180	2021-06-10 10:51:24.00000000	2025-06-04 10:33:11.00000000	364	0	6.2400000000000
13	Défaut Tension		1254	2021-06-20 18:02:15.00000000	2025-07-07 10:39:33.00000000	282	0	3.5400000000000
14	defaut communication avec A		1184	2021-09-20 13:47:06.00000000	2025-07-06 10:22:09.00000000	402	0	3.5600000000000
15	Niveau Très Bas		1038	2021-06-11 07:56:14.00000000	2025-05-14 21:13:19.00000000	303	0	3.0200000000000
16	Défaut communication avec B		1014	2021-09-22 20:50:43.00000000	2025-07-06 10:20:05.00000000	311	0	3.0600000000000
17	Etat communication		0	NULL	NULL	0	1	NULL
18	JUIN-JUILLET-AOUT		0	NULL	NULL	0	1	NULL
19	OCTOBER---->MARS		0	NULL	NULL	0	1	NULL
20	Pointe juin-juillet-aout		0	NULL	NULL	0	1	NULL
21	Pointe reste mois		0	NULL	NULL	0	1	NULL
22	SEPTEMBER		0	NULL	NULL	0	1	NULL
23	AVRIL-MAI		0	NULL	NULL	0	1	NULL
24	defaut modem		0	NULL	NULL	0	1	NULL
25	pointe avril-mai-sept		0	NULL	NULL	0	1	NULL

FIGURE 1 – Résultat d'exécution partiel : tous les indicateurs ne sont pas affichés

À la suite des analyses portant sur la fréquence, la complétude et la qualité des données, les variables suivantes ont été retenues pour la construction des vues analytiques principales :

- **TROZA_A** et **TROZA_B** : marche/arrêt, débit, production/24h.
- **Réservoir** : débit sortant 1 et 2, niveau d'eau, seuil très haut / très bas atteint.

Ces choix ont été motivés par la disponibilité des données, leur qualité mesurée lors de l'analyse, et leur utilité directe pour la modélisation prédictive du niveau d'eau par LSTM.

2. Analyse des écarts temporels entre les lectures

Listing 2 – Analyse des écarts temporels pour le site 1

```
WITH time_gaps AS (
    SELECT
        INF_NumberInStation,
        INF_Date,
        LAG(INF_Date) OVER (PARTITION BY INF_NumberInStation ORDER
            BY INF_Date) as prev_date,
        DATEDIFF(minute,
            LAG(INF_Date) OVER (PARTITION BY INF_NumberInStation
                ORDER BY INF_Date),
            INF_Date) as minutes_gap
    FROM dbo.View_ArchivedInformations
    WHERE STA_SiteNumber = 1
)
SELECT
    INF_NumberInStation,
    vi.INF_Label,
    AVG(CAST(minutes_gap as FLOAT)) as avg_minutes_between_readings
    ,
    MIN(minutes_gap) as min_gap_minutes,
    MAX(minutes_gap) as max_gap_minutes,
    COUNT(*) as gap_count
FROM time_gaps tg
LEFT JOIN dbo.View_Information vi
    ON tg.INF_NumberInStation = vi.INF_NumberInStation
    AND vi.STA_SiteNumber = 1
WHERE minutes_gap IS NOT NULL
GROUP BY INF_NumberInStation, vi.INF_Label
ORDER BY INF_NumberInStation;
```

L'exécution de cette requête a permis d'évaluer la fréquence de mise à jour des capteurs, en mesurant les écarts en minutes entre chaque lecture consécutive. Ces informations sont cruciales pour détecter des lectures irrégulières ou des lacunes dans les données.

	INF_NumberInStation	INF_Label	avg_minutes_between_readings	min_gap_minutes	max_gap_minutes	gap_count
1	1	ORDRE Marche/Arrêt	26.2615323473472	0	18584	70454
2	2	Débit	2.07790226449296	2	10050	884878
3	3	CTR Production	62.5877222865708	60	16920	29354
4	4	Production(/24h)	62.5877222865708	60	16920	29354
5	5	Production(/h)	62.5877222865708	60	16920	29354
6	6	Défaut Tension	1708.497700092	1	64223	1087
7	8	ETAT FORAGE	26.0710431021121	0	41618	70971
8	9	Mode Automatique	3160.28035714286	0	169607	560
9	10	Nbre heure	62.5877222865708	60	16920	29354
10	11	Nbre Dem	62.5877222865708	60	16920	29354
11	15	Mode Manuel	8590.10243902439	0	169607	205
12	16	Défaut Isolement	1710.07090239411	1	64223	1086
13	17	NIVEAU HAUT ARRET	86.8313565604151	0	152143	10792
14	18	NIVEAU BAS DEM	1.21990053498122	0	18598	1513698
15	6001	Défaut Equipement	229.100187382886	0	14558	8005

FIGURE 2 – Résultat d'exécution partiel : visualisation des écarts temporels

Sur la base des résultats, une agrégation temporelle à l'échelle horaire (1 heure) a été décidée afin d'homogénéiser les séries temporelles en vue de leur traitement par des modèles prédictifs.

Création des vues analytiques

Suite à l'analyse de la qualité et de la densité des données, nous avons procédé à la création de vues analytiques spécialisées pour chaque composant du système : les stations de pompage TROZA A et TROZA B, le réservoir, ainsi qu'une vue combinée. L'objectif était de produire des données propres, uniformisées et agrégées temporellement pour faciliter la modélisation prédictive.

Agrégation temporelle et fonctions utilisées

Toutes les vues ont été construites avec une **agrégation horaire**, c'est-à-dire que les données initialement collectées toutes les 2-30 minutes ont été regroupées par heure. Cette approche permet d'homogénéiser les séries temporelles et de réduire le bruit tout en conservant les tendances significatives.

- Les fonctions d'agrégation suivantes ont été appliquées selon la nature des variables :
- **AVG()** : pour les variables continues comme le débit et la production, afin d'obtenir une moyenne horaire représentative ;
 - **MAX()** : pour les variables de production cumulative sur 24h, prenant la valeur maximale dans l'heure ;
 - **Logique conditionnelle** : pour les variables binaires (Marche/Arrêt), dérivées des valeurs de débit (Marche si débit > 0).

Important : Aucune ligne contenant des valeurs NULL n'a pas été conservée dans les vues finales, garantissant ainsi la cohérence des données pour les algorithmes de machine learning.

1. Vue TROZA A (vw_TROZA_A)

Cette vue regroupe les données de la première station de pompage (STA_SiteNumber = 1) en se concentrant sur trois indicateurs clés :

- **Marche/Arrêt** : État binaire de la pompe, calculé automatiquement (1 si débit > 0, sinon 0)
- **Débit** : Débit de pompage en l/s, moyenné sur chaque heure
- **Production (24h)** : Volume produit sur 24h en m³, valeur maximale horaire

Listing 3 – Création de la vue TROZA A avec agrégation horaire

```
CREATE VIEW [dbo].[vw_TROZA_A] AS  
WITH RawData AS (
```

```

SELECT
    DATEADD(HOUR, DATEDIFF(HOUR, 0, ai.INF_Date), 0) AS Heure,
    inf.INF_Label,
    TRY_CAST(ai.INF_Value AS FLOAT) AS Valeur
FROM ScadaNetDb.dbo.View_ArchivedInformations ai
JOIN ScadaNetDb.dbo.View_Informations inf
    ON ai.STA_SiteNumber = inf.STA_SiteNumber
    AND ai.INF_NumberInStation = inf.INF_NumberInStation
WHERE ai.STA_SiteNumber = 1
    AND inf.INF_Label IN (
        'Marche/Arret',
        'D bit',
        'Production(/24h)'
    )
),
Aggregated AS (
    SELECT
        Heure,
        -- D rivation de Marche/Arr t depuis le D bit (1 =
        -- marche, 0 = arr t)
        CASE WHEN AVG(CASE WHEN INF_Label = 'D bit' THEN Valeur
            END) > 0
            THEN 1 ELSE 0 END AS Marche_Arret,
        AVG(CASE WHEN INF_Label = 'D bit' THEN Valeur END) AS
        Debit,
        AVG(CASE WHEN INF_Label = 'Production(/24h)' THEN Valeur
            END) AS Production_24h
    FROM RawData
    GROUP BY Heure
)
SELECT
    Heure,
    Marche_Arret AS [Marche/Arr t],
    Debit AS [D bit],
    Production_24h AS [Production (24h)]
FROM Aggregated;

```

	Heure	Marche/Arrêt	Débit	Production (24h)
1	2022-08-15 05:00:00.000	0	0	0
2	2022-08-29 10:00:00.000	0	0	0
3	2024-09-24 12:00:00.000	1	25.7813762029012	203
4	2022-03-02 10:00:00.000	1	6.8582818031311	1059
5	2024-01-04 15:00:00.000	1	15.8354314168294	0
6	2025-04-17 19:00:00.000	1	20.5479895273844	194
7	2022-10-11 19:00:00.000	1	23.4046693166097	0
8	2023-12-24 02:00:00.000	1	37.1871517817179	0
9	2023-03-14 22:00:00.000	1	24.0405127207438	1123
10	2024-06-10 11:00:00.000	1	28.3075455347697	1802
11	2025-04-23 07:00:00.000	1	25.3998243967692	666
12	2025-05-10 06:00:00.000	1	15.131075100104	838
13	2023-07-14 14:00:00.000	1	31.0948729832967	0
14	2025-03-21 14:00:00.000	1	27.1515600840251	960
15	2022-06-01 02:00:00.000	1	25.1631570180257	391
16	2024-03-31 14:00:00.000	1	27.6909285227458	1220
17	2024-04-22 00:00:00.000	1	22.0271105500014	655

FIGURE 3 – Résultat d'exécution de la vue TROZA A - Données agrégées par heure

2. Vue TROZA B (vw_TROZA_B)

Cette vue traite les données de la seconde station de pompage (STA_SiteNumber=2) avec la même logique d'agrégation, mais adaptée aux spécificités de cette station :

Listing 4 – Création de la vue TROZA B avec gestion des valeurs nulles

```
CREATE VIEW [dbo].[vw_TROZA_B] AS
WITH RawData_B AS (
    SELECT
        DATEADD(HOUR, DATEDIFF(HOUR, 0, ai.INF_Date), 0) AS Heure,
        inf.INF_Label,
        TRY_CAST(ai.INF_Value AS FLOAT) AS Valeur,
        inf.INF_IsLogical
    FROM ScadaNetDb.dbo.View_ArchivedInformations ai
    JOIN ScadaNetDb.dbo.View_Informations inf
        ON ai.STA_SiteNumber = inf.STA_SiteNumber
        AND ai.INF_NumberInStation = inf.INF_NumberInStation
    WHERE ai.STA_SiteNumber = 2
        AND inf.INF_Label IN (
            'Marche/Arret',
            'D bit',
            'Production(/24h)'
        )
),
Aggregated_B AS (
    SELECT
        Heure,
        ISNULL(AVG(CASE WHEN INF_Label = 'D bit' THEN Valeur END),
            0) AS Debit,
        ISNULL(MAX(CASE WHEN INF_Label = 'Production(/24h)' THEN
            Valeur END), 0) AS Production_24h
    FROM RawData_B
    GROUP BY Heure
),
FinalOutput AS (
```

```

SELECT
    Heure ,
    -- Logique Marche/Arr t : 0 si D bit=0, sinon 1
    CASE WHEN Debit = 0 THEN CAST(0 AS BIT) ELSE CAST(1 AS BIT)
    END AS Marche_Arret ,
    Debit ,
    Production_24h
FROM Aggregated_B
)
SELECT
    f.Heure ,
    f.Marche_Arret AS [Marche/Arr t],
    f.Debit AS [D bit (l/s)],
    f.Production_24h AS [Production (/24h) (m )]
FROM FinalOutput f;

```

	Heure	Marche/Arrêt	Débit (l/s)	Production (/24h) (m³)
1	2025-07-09 10:00:00.000	1	16,3772309573198	1255
2	2025-07-09 09:00:00.000	1	18,0352228164673	1255
3	2025-07-09 08:00:00.000	1	18,0049592336019	1255
4	2025-07-09 07:00:00.000	1	17,4259048059583	1255
5	2025-07-09 06:00:00.000	1	17,5183488443494	1168
6	2025-07-09 05:00:00.000	1	18,0505078633626	1168
7	2025-07-09 04:00:00.000	1	18,1044737497966	1168
8	2025-07-09 03:00:00.000	1	17,8594140052795	1168
9	2025-07-09 02:00:00.000	1	18,0717434565226	1168
10	2025-07-09 01:00:00.000	1	17,4670788757503	1168
11	2025-07-09 00:00:00.000	1	18,0451154708862	1168
12	2025-07-08 23:00:00.000	1	18,0583913167318	1168
13	2025-07-08 22:00:00.000	1	18,0808987299601	1168
14	2025-07-08 21:00:00.000	1	15,6964729761084	1168
15	2025-07-08 20:00:00.000	1	17,4480048172176	1168
16	2025-07-08 19:00:00.000	1	17,5039548548559	1168

FIGURE 4 – Résultat d'exécution de la vue TROZA B - Station de pompage 2

3. Vue Réservoir (View_Reservoir)

Cette vue centralise les données du réservoir (STA_SiteNumber = 3) en se concentrant sur le niveau d'eau et les débits de distribution. Elle intègre également des seuils critiques pour détecter les niveaux très hauts et très bas :

- **NIVEAU** : Niveau d'eau dans le réservoir, moyenné par heure
- **Débit Dist DN250/DN100** : Débits de distribution sur les conduites de différents diamètres, moyenné par heure
- **Seuils critiques** : Indicateurs binaires pour niveau très haut (2.008 m) et très bas (0.179 m)

Listing 5 – Vue réservoir avec calcul des seuils critiques

```

CREATE VIEW [dbo].[View_Reservoir] AS
WITH hourly_data AS (
    SELECT

```



```

        DATEADD(HOUR, DATEDIFF(HOUR, 0, vai.INF_Date), 0) as
            datetime_hour,
        vi.INF_Label,
        AVG(vai.INF_Value) as avg_value
FROM dbo.View_Information vi
JOIN dbo.View_ArchivedInformations vai
    ON vi.STA_SiteNumber = vai.STA_SiteNumber
    AND vi.INF_NumberInStation = vai.INF_NumberInStation
WHERE vi.STA_SiteNumber = 3
    AND vi.INF_Label IN ('NIVEAU', 'D bit_Dist_DN250', 'D bit
        _Dist_DN100')
    AND vai.INF_Value IS NOT NULL
GROUP BY DATEADD(HOUR, DATEDIFF(HOUR, 0, vai.INF_Date), 0), vi.
    INF_Label
)
SELECT
    datetime_hour,
    MAX(CASE WHEN INF_Label = 'NIVEAU' THEN avg_value END) as
        niveau_avg,
    MAX(CASE WHEN INF_Label = 'D bit_Dist_DN250' THEN avg_value
        END) as debit_dn250_avg,
    MAX(CASE WHEN INF_Label = 'D bit_Dist_DN100' THEN avg_value
        END) as debit_dn100_avg,

    -- Seuil tr s haut ( 2.008 m)
    CASE
        WHEN MAX(CASE WHEN INF_Label = 'NIVEAU' THEN avg_value END)
            >= 2.00786304473877 THEN 1
        ELSE 0
    END as niveau_tres_haut,

    -- Seuil tr s bas ( 0.179 m)
    CASE
        WHEN MAX(CASE WHEN INF_Label = 'NIVEAU' THEN avg_value END)
            <= 0.179144926309585 THEN 1
        ELSE 0
    END as niveau_tres_bas
FROM hourly_data
GROUP BY datetime_hour
ORDER BY datetime_hour;

```

	datetime_hour	niveau_avg	debit_dn250_avg	debit_dn100_avg	niveau_tres_haut	niveau_tres_bas
14	2021-06-10 23:00:00.000	1,71421331564585	28,8709595362345	0	0	0
15	2021-06-11 00:00:00.000	1,73685357968012	28,9379721959432	0	0	0
16	2021-06-11 01:00:00.000	1,99209327697754	27,7026270548503	0	0	0
17	2021-06-11 02:00:00.000	1,61180860201518	27,6637166341146	0	0	0
18	2021-06-11 03:00:00.000	1,95321302811305	26,2587422053019	0	0	0
19	2021-06-11 04:00:00.000	1,78785239855448	24,7508963267008	0	0	0
20	2021-06-11 05:00:00.000	2,18756257692973	25,1107549031576	3,43503980636597	1	0
21	2021-06-11 06:00:00.000	1,75725007851919	26,2614124298096	12,8612650235494	0	0
22	2021-06-11 07:00:00.000	0,911818009614...	23,5308107058207	16,0728161493937	0	0
23	2021-06-11 08:00:00.000	0,657809002200...	57,6063680330912	18,8111850102743	0	0
24	2021-06-11 09:00:00.000	1,73051763772964	26,8240891138713	17,8868238449097	0	0
25	2021-06-11 10:00:00.000	1,00616474350293	25,9735255559286	18,697571182251	0	0
26	2021-06-11 11:00:00.000	1,75263848702113	25,7376466751099	16,3110094388326	0	0
27	2021-06-11 12:00:00.000	1,41926863988241	20,5378805796305	7,40635277430216	0	0
28	2021-06-11 13:00:00.000	1,83051245212555	12,8154167493184	4,22880392074585	0	0
29	2021-06-11 14:00:00.000	1,76774603525798	19,1461308479309	3,85317767461141	0	0

FIGURE 5 – Résultat d'exécution de la vue Réservoir

4. Vue combinée (intégrant toutes les données)

Une vue finale combinée a été créée en joignant les vues TROZA A, TROZA B, et Réservoir sur la colonne horaire commune, fournissant ainsi un tableau consolidé pour l'analyse globale.

Listing 6 – Vue combinée des données des stations et réservoir

```
CREATE VIEW [dbo].[vw_Combined] AS
SELECT
    a.Heure,
    a.[Marche/Arrêt] AS Marche_Arret_A,
    a.[Débit] AS Debit_A,
    a.[Production (24h)] AS Production_24h_A,
    b.[Marche/Arrêt] AS Marche_Arret_B,
    b.[Débit (l/s)] AS Debit_B,
    b.[Production (/24h) (m³)] AS Production_24h_B,
    r.niveau_avg,
    r.debit_dn250_avg,
    r.debit_dn100_avg,
    r.niveau_tres_haut,
    r.niveau_tres_bas
FROM dbo.vw_TROZA_A a
LEFT JOIN dbo.vw_TROZA_B b ON a.Heure = b.Heure
LEFT JOIN dbo.View_Reservoir r ON a.Heure = r.datetime_hour
ORDER BY a.Heure;
```

Ces vues permettent d'avoir des données propres, agrégées, prêtes pour l'utilisation dans des modèles de prévision basés sur l'IA.

- **Horodatage uniforme** : Agrégation horaire cohérente sur toutes les vues
- **Qualité des données** : Élimination des valeurs NULL et des incohérences
- **Variables dérivées** : Calcul automatique de l'état Marche/Arrêt et des seuils critiques
- **Cohérence temporelle** : Synchronisation des timestamps entre les différentes sources

Conclusion

Ce travail de préparation des données SCADA constitue une étape fondamentale et critique du projet de prévision du niveau d'eau par intelligence artificielle. L'analyse approfondie menée sur les données de la station de Kairouan a permis d'identifier et de résoudre les principales problématiques liées à la qualité des données : incohérences temporelles, valeurs manquantes, et nomenclatures ambiguës.

La création des vues analytiques spécialisées (`vw_TROZA_A`, `vw_TROZA_B`, `View_Reservoir`, et `vw_Combined`) avec une agrégation horaire uniforme garantit désormais la disponibilité de données propres, cohérentes et exploitables pour les algorithmes de machine learning. Cette phase de traitement BI, bien qu'exigeante en temps et en ressources, **représente la partie la plus chronophage du projet**, nécessitant une attention particulière à chaque étape pour assurer la fiabilité des analyses futures.

Prochaines étapes :

Fort de ces données structurées et validées, la prochaine phase consistera à développer le modèle de prévision en utilisant **Python** comme langage de programmation principal. L'implémentation s'appuiera sur des algorithmes de réseaux de neurones récurrents (LSTM) pour prédire l'évolution du niveau d'eau dans le réservoir.

L'investissement en temps consacré à la préparation des données garantit une base solide pour des prédictions fiables et précises, condition essentielle au succès du déploiement du système de prévision par IA chez I3T.