# Etude de la régularité des trains entre 2015 et 2020 (SNCF)

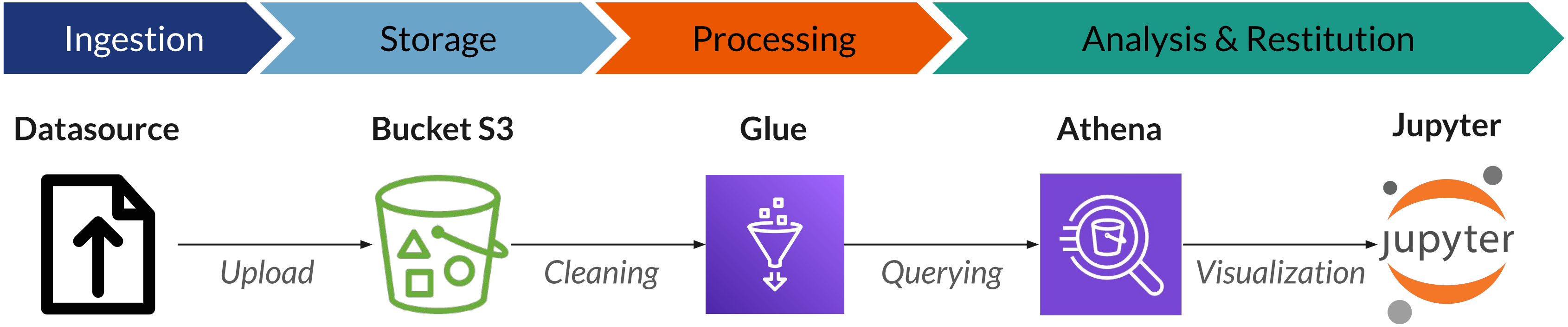
## Projet - Pipeline de traitements de données pour le cloud

CHAOUL Marc et TAURAND Clément, ING3 FISA
EPISEN SI Ing3

# Les 5 V

- **Volume :** Historique complet de 2015 à 2020 (cumul par mois sur les trajets communiqués par la SNCF). *7806 lignes soit 7806/(12*6) = 108 trajets en moyenne par mois*
- **Vélocité :** Données mensuelles. Le pipeline est conçu en mode "Batch" (traitement par lots à l'arrivée de nouveaux fichiers mensuels).
- **Variété :** Données structurées (CSV) avec schéma mixte (texte, numériques, pourcentages).
- **Véracité :** Données issues de l'Open Data SNCF (source officielle et fiable).
- **Valeur :** Prédiction des risques de retard selon la durée, transparence pour les usagers.

# Diagramme d’architecture



Ingestion

Storage

Processing

Analysis & Restitution

Datasource

Bucket S3

Glue

Athena

Jupyter

*Upload*

*Cleaning*

*Querying*

*Visualization*

Téléchargement des données (csv) depuis Kaggle

Nettoyage, conversion en parquet, légers traitements

Sélection des données et calculs

Affichage des résultats

# Format des Données → Raw Data

```
{
  "Year": 2019,
  "Month": 7.0,
  "Departure station": "ANGOULEME",
  "Arrival station": "PARIS MONTPARNASSE",
  "Average travel time (min)": 131.914979757,
  "Number of expected circulations": 247.0,
  "Number of cancelled trains": 0.0,
  "Number of late trains at departure": 191.0,
  "Average delay of late departing trains (min)":
3.5763525305400004,
  "Average delay of all departing trains (min)":
2.67827260459,
  "Comment (optional) delays at departure": null,
  "Number of trains late on arrival": 41.0,
  "Average delay of late arriving trains (min)":
22.924796748000002,
  "Average delay of all arriving trains (min)":
5.23333333333,
  "Comment (optional) delays on arrival": null,
  "% trains late due to external causes (weather,
obstacles, suspicious packages, malevolence, social
movements, etc.)": 0.25,
…
```

```
…
"% trains late due to railway infrastructure (maintenance,
works)": 0.15,
  "% trains late due to traffic management (rail line traffic,
network interactions)": 0.275,
  "% trains late due to rolling stock": 0.125,
  "% trains late due to station management and reuse of
material": 0.025,
  "% trains late due to passenger traffic (affluence, PSH
management, connections)": 0.175,
  "Number of late trains > 15min": 21.0,
  "Average train delay > 15min": 32.9658730159,
  "Number of late trains > 30min": 7.0,
  "Number of late trains > 60min": 2.0,
  "Period": "2019-07",
  "Delay due to external causes": 25.0,
  "Delay due to railway infrastructure": 15.0,
  "Delay due to traffic management": 27.500000000000004,
  "Delay due to rolling stock": 12.5,
  "Delay due to station management and reuse of material": 2.5,
  "Delay due to travellers taken into account": 17.5
}
```

# Format des Données → Cleaned Data

Données Supprimées

```
{
  "Year": 2019,
  "Month": 7.0,
  "Departure station": "ANGOULEME",
  "Arrival station": "PARIS MONTPARNASSE",
  "Average travel time (min)": 131.914979757,
  "Number of expected circulations": 247.0,
  "Number of cancelled trains": 0.0,
  "Number of late trains at departure": 191.0,
  "Average delay of late departing trains (min)":
3.5763525305400004,
  "Average delay of all departing trains (min)":
2.67827260459,
  "Comment (optional) delays at departure": null,
  "Number of trains late on arrival": 41.0,
  "Average delay of late arriving trains (min)":
22.924796748000002,
  "Average delay of all arriving trains (min)":
5.23333333333,
  "Comment (optional) delays on arrival": null,
  "% trains late due to external causes (weather,
obstacles, suspicious packages, malevolence, social
movements, etc.)": 0.25,
…
```

```
…
"% trains late due to railway infrastructure (maintenance,
works)": 0.15,
  "% trains late due to traffic management (rail line traffic,
network interactions)": 0.275,
  "% trains late due to rolling stock": 0.125,
  "% trains late due to station management and reuse of
material": 0.025,
  "% trains late due to passenger traffic (affluence, PSH
management, connections)": 0.175,
  "Number of late trains > 15min": 21.0,
  "Average train delay > 15min": 32.9658730159,
  "Number of late trains > 30min": 7.0,
  "Number of late trains > 60min": 2.0,
  "Period": "2019-07",
  "Delay due to external causes": 25.0,
  "Delay due to railway infrastructure": 15.0,
  "Delay due to traffic management": 27.500000000000004,
  "Delay due to rolling stock": 12.5,
  "Delay due to station management and reuse of material": 2.5,
  "Delay due to travellers taken into account": 17.5
}
```

# Format des Données → Standardized Data

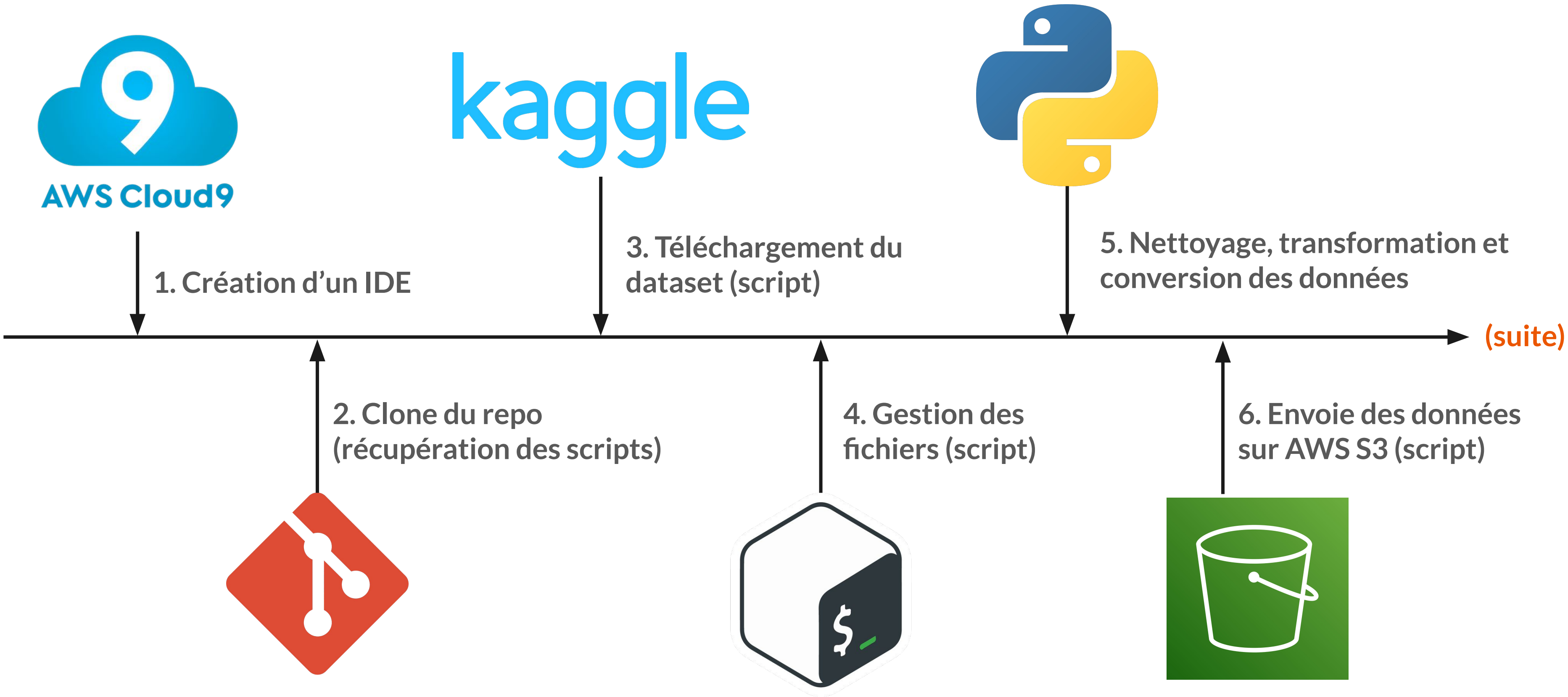
Ancien nom colonne
Nouveau nom colonne
Nouvelle colonne

```
{
  "Year": 2019,
  "Month": 7.0,
  "Departure station": "ANGOULEME",
  "Arrival station": "PARIS MONTPARNASSE",
  "Average travel time (min)": 131.914979757,
  "Number of expected circulations": 247.0,
  "Number of late trains at departure": 191.0,
  "Number of trains late on arrival": 41.0,
  "Number of late trains > 15min": 21.0,
  "Number of late trains > 30min": 7.0,
  "Number of late trains > 60min": 2.0,
  "Period": "2019-07",
  "Delay due to external causes": 25.0,
  "Delay due to railway infrastructure": 15.0,
  "Delay due to traffic management":
27.500000000000004,
  "Delay due to rolling stock": 12.5,
  "Delay due to station management and reuse of
material": 2.5,
  "Delay due to travellers taken into account": 17.5
}
```

→

```
{
  "year": 2019,
  "month": 7.0,
  "departure_station": "ANGOULEME",
  "arrival_station": "PARIS MONTPARNASSE",
  "avg_travel_time_min": 131.915,
  "nb_expected": 247.0, // this - nb_cancelled
  "nb_late_dep": 191.0,
  "nb_late_arr": 41.0,
  "nb_late_over_15": 21.0,
  "nb_late_over_30": 7.0,
  "nb_late_over_60": 2.0,
  "period": "2019-07",
  "delay_cause_external": 25.0,
  "delay_cause_infra": 15.0,
  "delay_cause_traffic": 27.5,
  "delay_cause_rolling_stock": 12.5,
  "delay_cause_station": 2.5,
  "delay_cause_travelers": 17.5,
  "nb_late_before_15": 20.0 //nb_late_arr - nb_late_over_15
}
```

# Étapes de construction et exécution de la pipeline



1. Création d'un IDE

2. Clone du repo (récupération des scripts)

3. Téléchargement du dataset (script)

4. Gestion des fichiers (script)

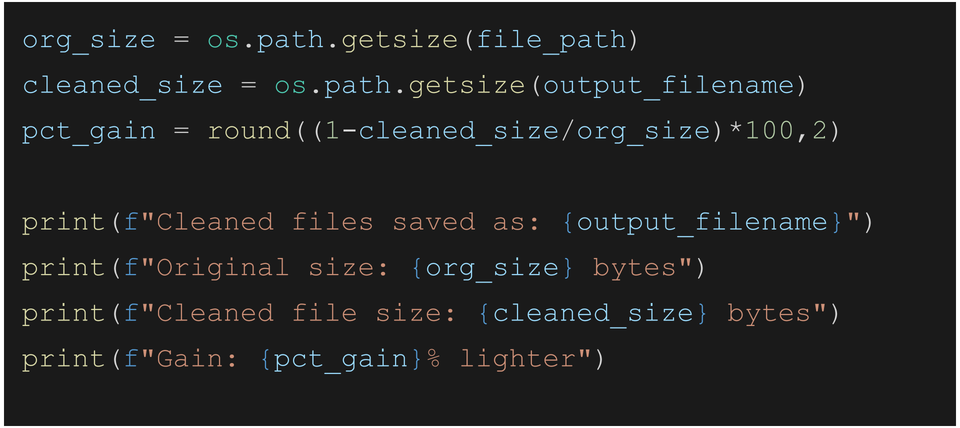
5. Nettoyage, transformation et conversion des données

6. Envoie des données sur AWS S3 (script)

(suite)

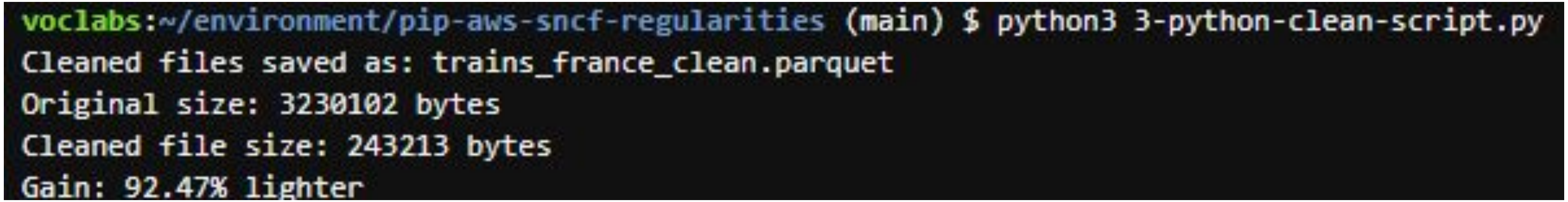
# Zoom : Nettoyage des données et conversion

CSV
+ Suppression de 45% des données
+ conversion en PARQUET
→ 92% plus léger



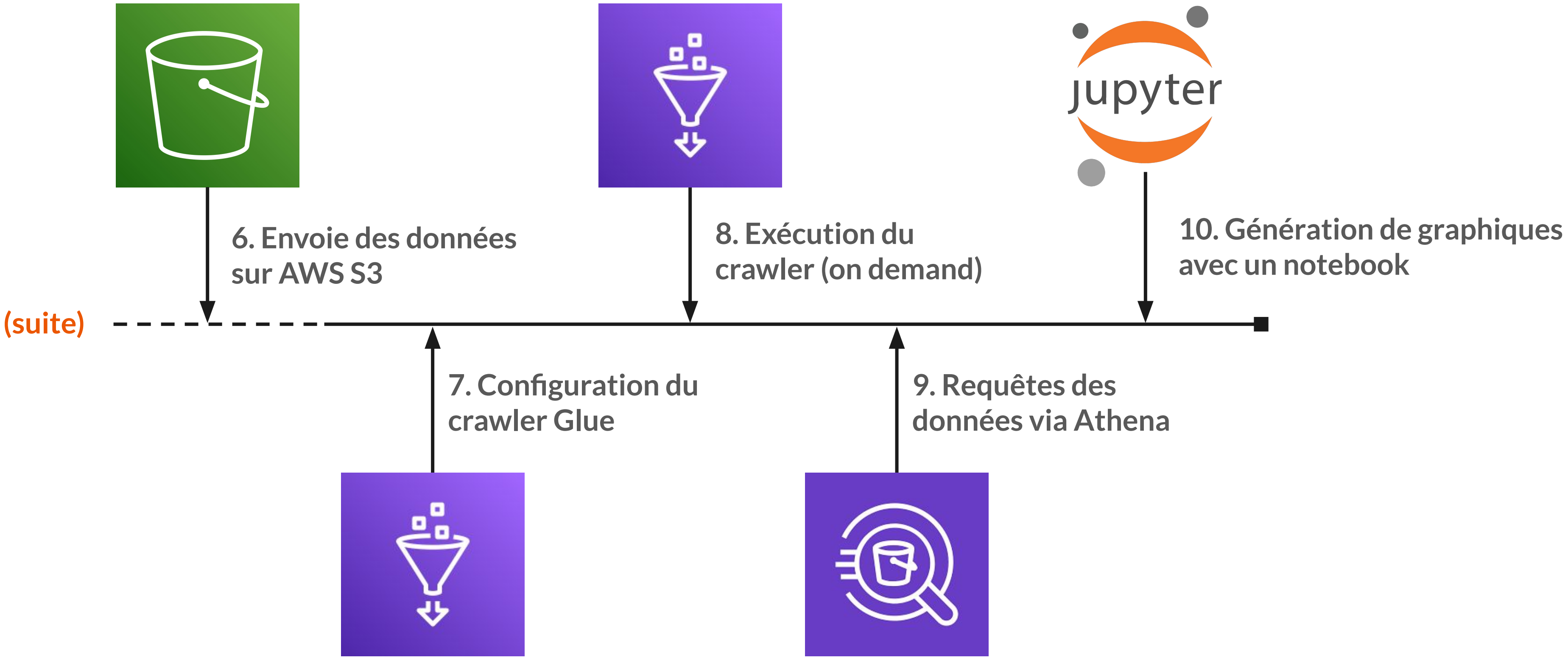
```
org_size = os.path.getsize(file_path)
cleaned_size = os.path.getsize(output_filename)
pct_gain = round((1-cleaned_size/org_size)*100,2)

print(f"Cleaned files saved as: {output_filename}")
print(f"Original size: {org_size} bytes")
print(f"Cleaned file size: {cleaned_size} bytes")
print(f"Gain: {pct_gain}% lighter")
```



```
voclabs:~/environment/pip-aws-sncf-regularities (main) $ python3 3-python-clean-script.py
Cleaned files saved as: trains_france_clean.parquet
Original size: 3230102 bytes
Cleaned file size: 243213 bytes
Gain: 92.47% lighter
```

# Étapes de construction et exécution de la pipeline



6. Envoie des données sur AWS S3

7. Configuration du crawler Glue

8. Exécution du crawler (on demand)

9. Requêtes des données via Athena

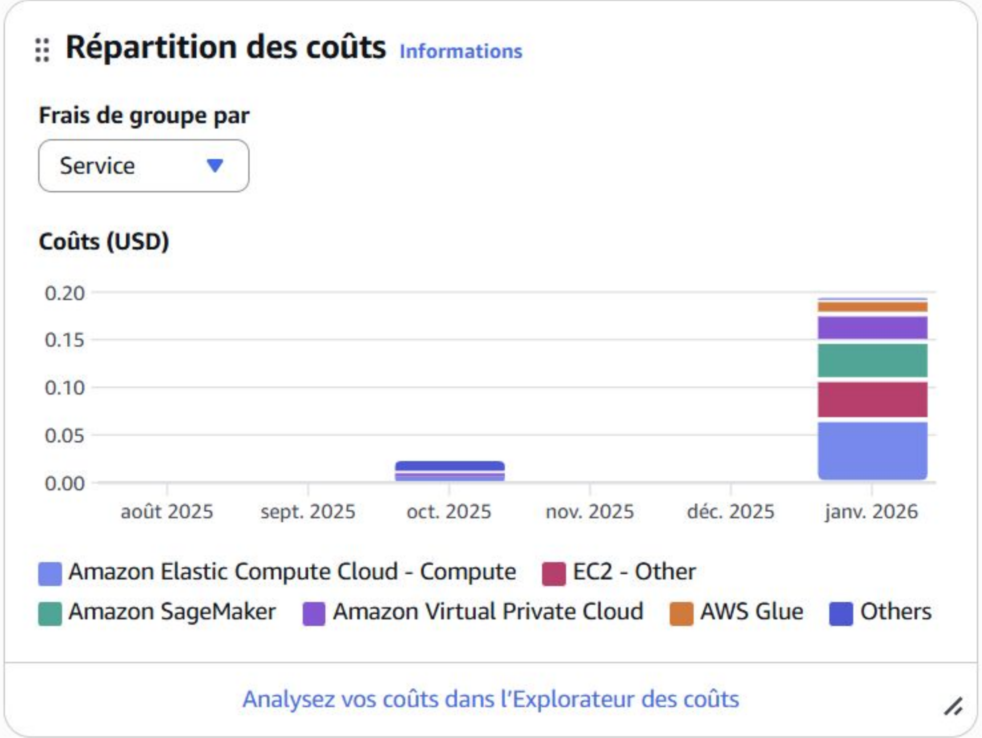
10. Génération de graphiques avec un notebook

# Les KPIs

1. Distribution des raisons de délais (retards) en fonction du nombre totale de retards à l'arrivée, groupé par période
2. Distribution des horaires de retards en fonction du nombre totale de retards à l'arrivée, groupé par période
3. Comparaison du nombre de train en retard par rapport au nombre de train prévus
4. Comparaison des lignes ayant le plus de retard (cumul)

# Estimation des coûts

NB :
EC2 et SageMaker ne sont pas utilisés dans la pipeline (simplement testés lors de la construction), donc les coûts sont encore moins conséquents



Répartition des coûts Informations

Frais de groupe par: Service

Coûts (USD)

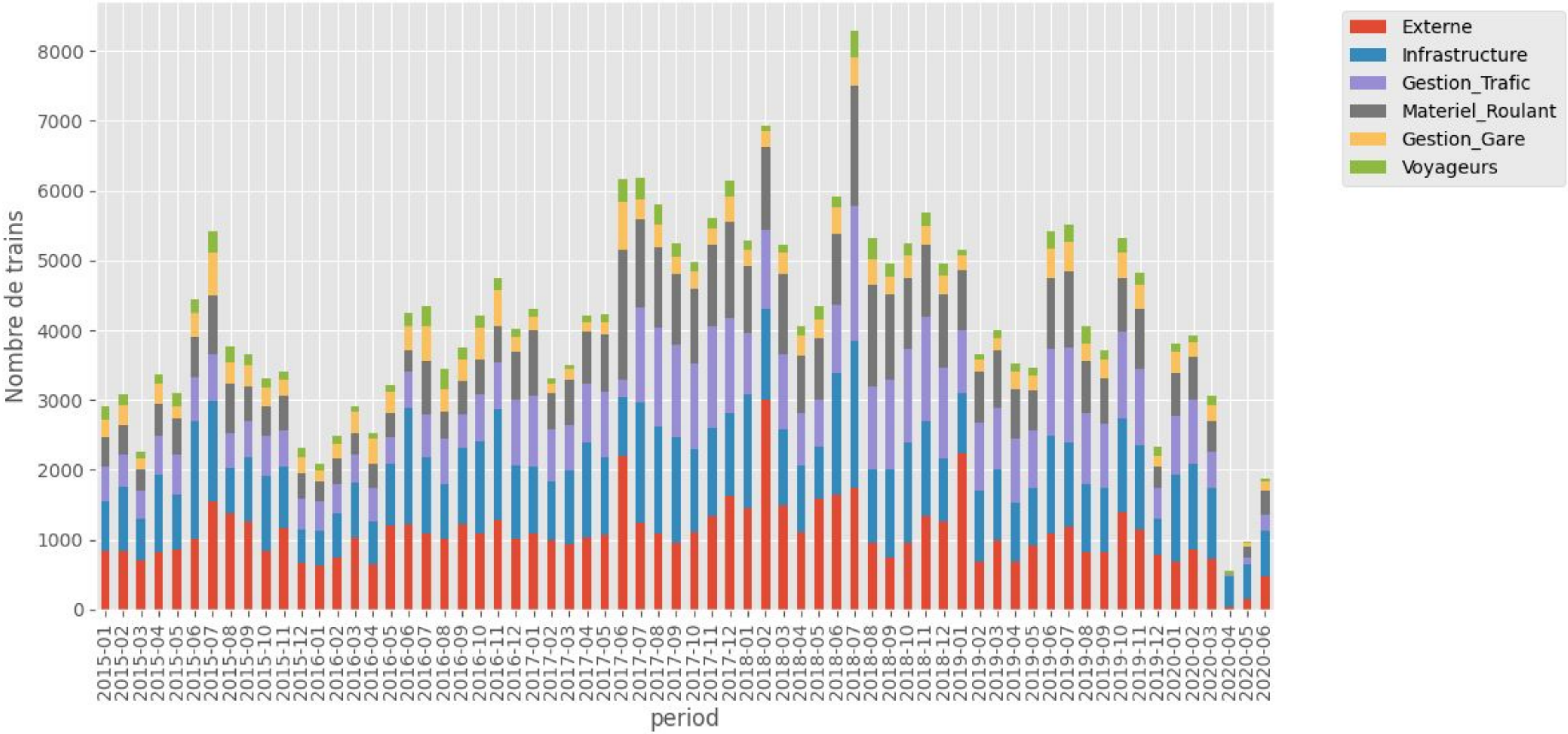
Analysez vos coûts dans l'Explorateur des coûts

# Ressources

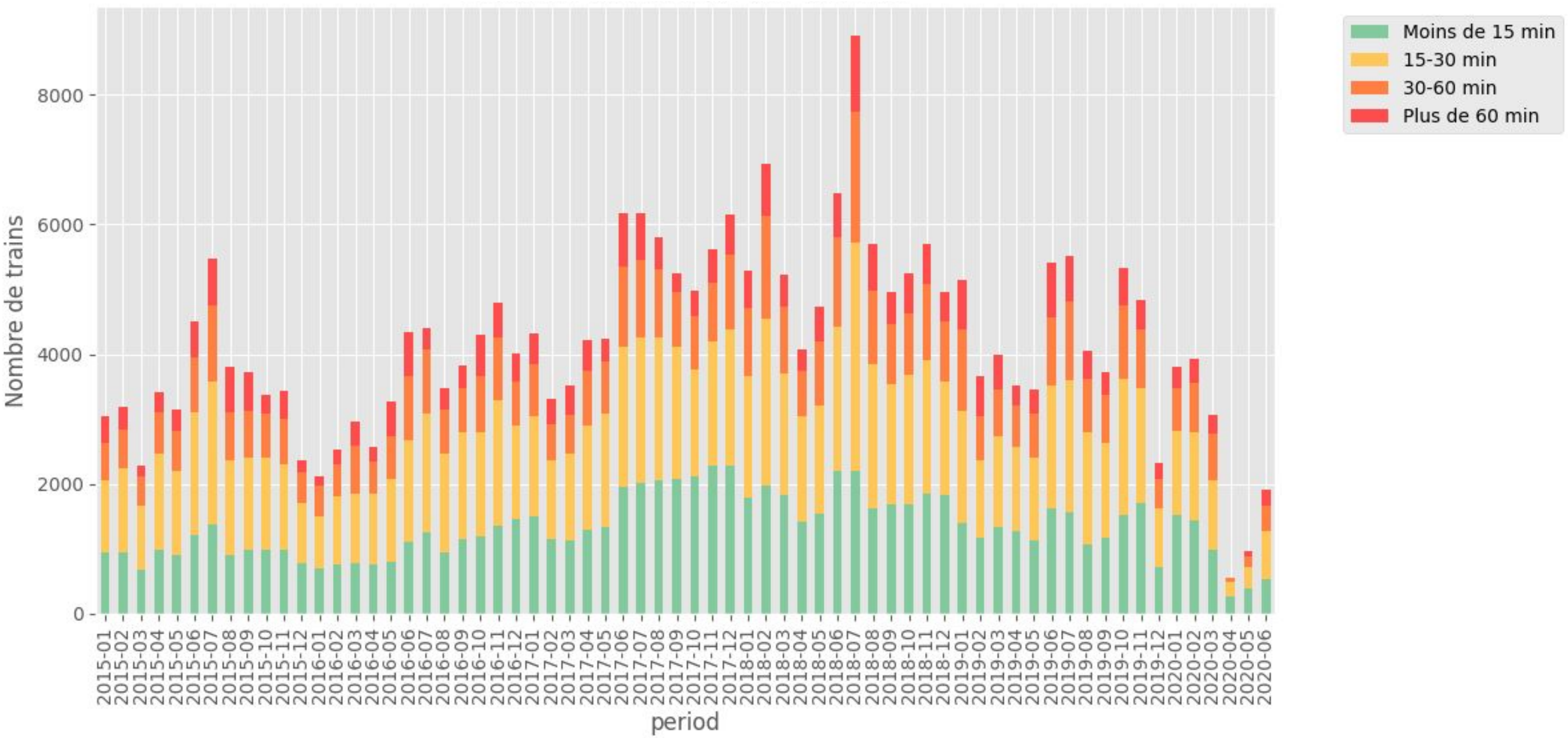
Repo GitHub :
https://github.com/BlooSkyd/pip-aws-sncf-regularities

Dataset :
https://www.kaggle.com/datasets/gatandubuc/public-transport-traffic-data-in-france

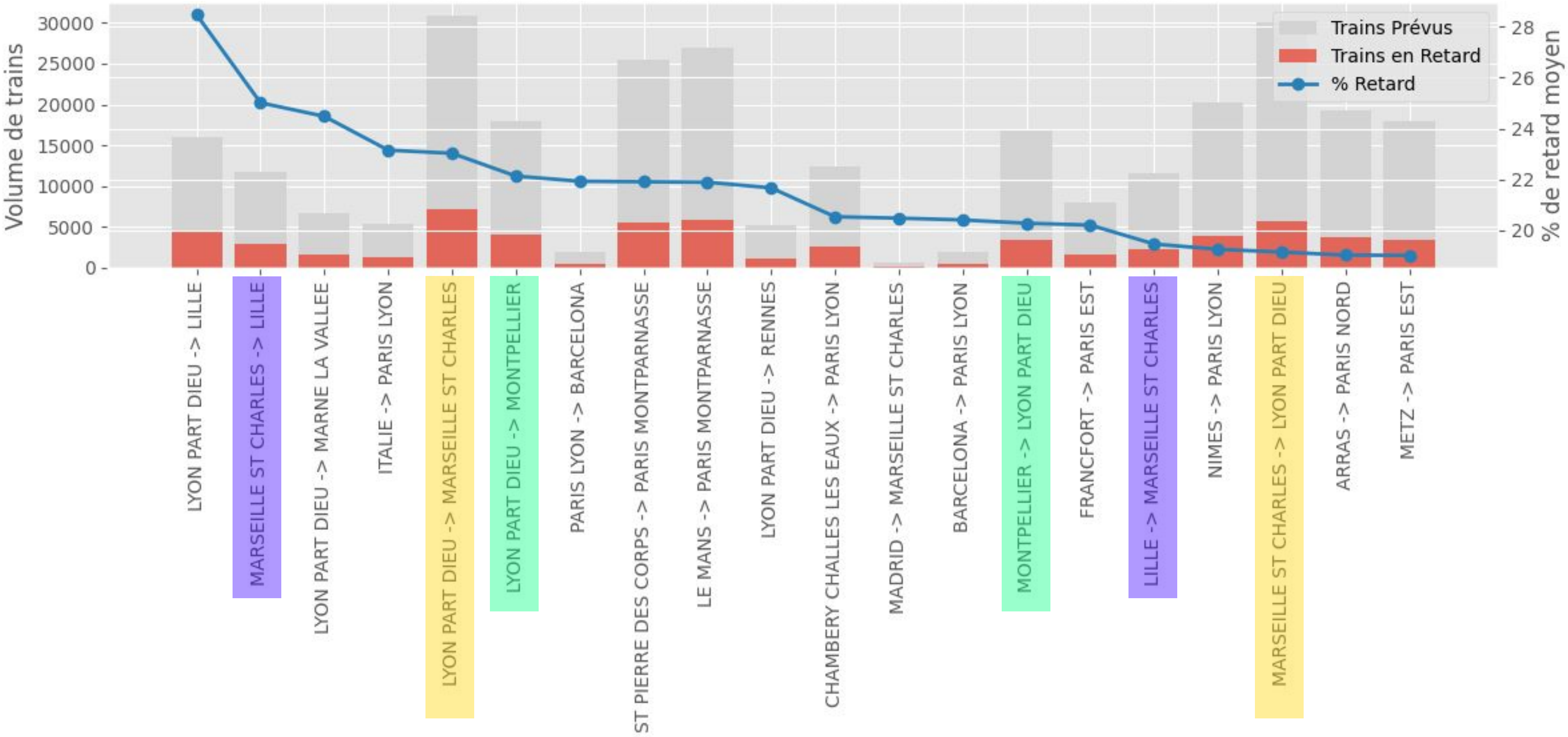
# Distribution des causes de retards par période



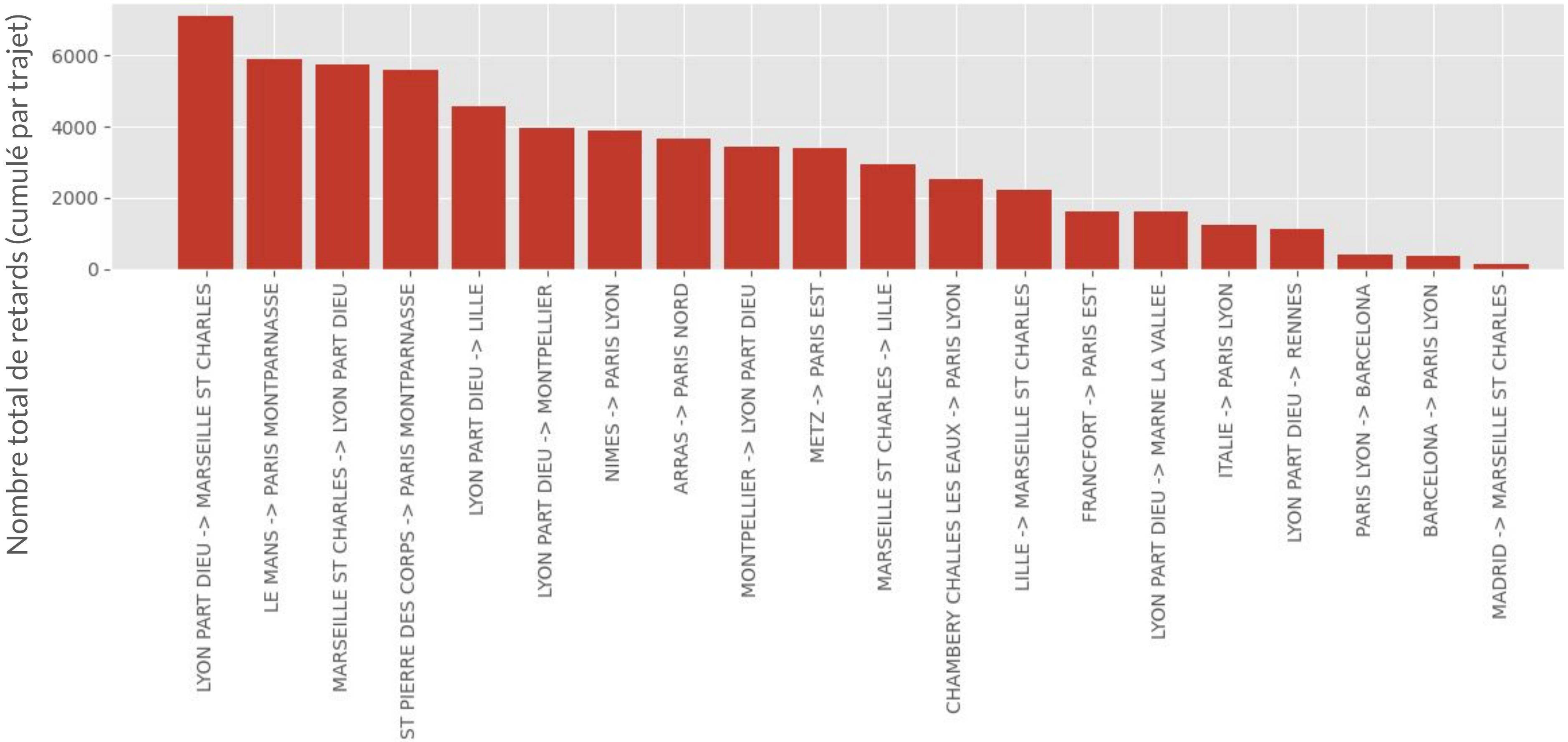
# Sévérité des retards par période



# Comparaison Volume vs. Taux de retard (top 20)



# Lignes les plus problématiques (en volume, top 20)


Nombre total de retards (cumulé par trajet)
LYON PART DIEU -> MARSEILLE ST CHARLES
LE MANS -> PARIS MONTPARNASSE
MARSEILLE ST CHARLES -> LYON PART DIEU
ST PIERRE DES CORPS -> PARIS MONTPARNASSE
LYON PART DIEU -> LILLE
LYON PART DIEU -> MONTPELLIER
NIMES -> PARIS LYON
ARRAS -> PARIS NORD
MONTPELLIER -> LYON PART DIEU
METZ -> PARIS EST
MARSEILLE ST CHARLES -> LILLE
CHAMBERY CHALLES LES EAUX -> PARIS LYON
LILLE -> MARSEILLE ST CHARLES
FRANCFORT -> PARIS EST
LYON PART DIEU -> MARNE LA VALLEE
ITALIE -> PARIS LYON
LYON PART DIEU -> RENNES
PARIS LYON -> BARCELONA
BARCELONA -> PARIS LYON
MADRID -> MARSEILLE ST CHARLES