enigma

Ecole du Numérique, Informatique, Global data et Management Appliqués

PROJET
ELK
SNCF

Jean-Gilles Kyllian **2025**

INTRODUCTION

Dans un contexte où la mobilité durable et intelligente devient un enjeu majeur pour les grandes métropoles, la SNCF met à disposition de nombreux jeux de données ouvertes permettant d'analyser le trafic ferroviaire en France. Ces données constituent une source précieuse pour comprendre les dynamiques de déplacement, anticiper les flux voyageurs et améliorer la qualité de service.

Le thème choisi « Mobilité Voyageurs » vise à concevoir un système automatisé de collecte, d'analyse et de visualisation de données ferroviaires en temps réel.

L'objectif principal est de suivre la fréquentation et les départs des trains depuis la gare de Lille Flandres, en combinant deux types de données :

- Les données historiques issues du fichier frequentation-gares.csv, fournissant des indicateurs de trafic sur plusieurs années ;

- Les données temps réel provenant de l'API publique de la SNCF, renseignent les départs actualisés des trains toutes les quelques minutes.

Pour traiter ces informations, le projet s'appuie sur la stack ELK (Elasticsearch, Logstash, Kibana) un écosystème open source dédié à la gestion, l'indexation et la visualisation de données à grande échelle.

La pipeline développé en Python automatise l'extraction, le nettoyage et l'injection des données dans Elastic Cloud, avant leur exploitation dans Kibana sous forme de dashboards interactifs.

Ce travail vise ainsi à démontrer la faisabilité d'une surveillance en temps réel de la mobilité ferroviaire à l'échelle locale, tout en valorisant les compétences techniques liées à l'intégration de données, à l'automatisation des flux et à la visualisation analytique.

MÉTHODOLOGIE

La démarche adoptée pour ce projet s'appuie sur une approche data engineering complète, allant de la collecte des données à leur visualisation dans un environnement Elastic Cloud. Elle se décompose en quatre grandes étapes : l**a collecte, le traitement, l'indexation et la visualisation.**

1. Collecte des données

Deux sources de données ont été exploitées afin de couvrir à la fois la vision historique et la vision temps réel de la mobilité ferroviaire :

Les données historiques proviennent du fichier CSV **frequentation-gares.csv** publié par la SNCF, contenant la fréquentation annuelle de plus de 3 000 gares françaises.

- Les données en temps réel sont issues de l'API publique **https://api.sncf.com/v1/coverage/sncf/stop_areas/.../departures**, qui permet d'obtenir à tout moment les prochains départs depuis une gare donnée.

- Dans le cadre de ce projet, l'analyse s'est concentrée sur la gare de Lille Flandres, identifiée par son code SNCF **stop_area:SNCF:87286005.**

2. Traitement et nettoyage des données

Un script Python a été développé pour automatiser la récupération et la préparation des données.

Les principales étapes de traitement incluent :

- La suppression des colonnes non pertinentes (codes UIC, segmentation, etc.)

- La conversion des valeurs manquantes en None afin d'éviter les erreurs d'indexation (NaN non supporté par Elasticsearch) ;

- L'ajout d'un champ timestamp pour chaque enregistrement, permettant de suivre l'évolution temporelle des données ;

- La génération d'un identifiant unique pour chaque document, afin d'éviter les doublons lors des réindexations.

3. Indexation dans Elasticsearch

Les données nettoyées sont ensuite envoyées vers Elastic Cloud, à l'aide du module Python elasticsearch et de la méthode helpers.bulk() pour une indexation en masse.

Deux index distincts sont créés :

- **historique** : contenant les statistiques annuelles de fréquentation des gares ;
- **sncf_lille_realtime** : alimenté toutes les 15 minutes avec les départs récents de Lille Flandres.

Les identifiants de connexion (cloud_id, user, password) sont stockés dans un fichier .env (normalement on ne partage pas sur github en le mettant dans le .gitignore mais pour les tests je l'ai laissé), garantissant la sécurité et la portabilité du projet.

4. Visualisation et analyse avec Kibana

Une fois les données indexées, l'outil Kibana est utilisé pour construire un dashboard interactif.

Ce tableau de bord permet de visualiser :

- Les départs en temps réel (horaires, destinations, types de trains) ;
- Les volumes de fréquentation par gare ;
- Les tendances horaires et les pics d'activité ;
- Les indicateurs clés (KPI) sur la mobilité régionale.
- Cette visualisation permet une compréhension immédiate des dynamiques de transport à Lille et facilite la prise de décision fondée sur les données.

INSIGHT

Les visualisations réalisées sur Kibana offrent une lecture claire et hiérarchisée de la fréquentation des gares françaises, en combinant données historiques 2024 et temps réel (départs à Lille Flandres).

Les gares les plus fréquentées en France

Le premier graphique montre que la Gare du Nord (Paris) domine très largement avec environ 260 millions de voyageurs en 2024, suivie par Paris Gare de Lyon et Paris Est, autour de 110 millions. Cette concentration des flux sur la région parisienne illustre la centralisation du réseau ferroviaire français. Ces volumes massifs confirment que Paris agit comme hub national, et justifient la nécessité d'une gestion fine de la capacité et de la maintenance.

Focus sur la Métropole Européenne de Lille (MEL)

Le graphique des gares les plus fréquentées de la MEL met en évidence la prédominance de Lille Flandres (≈ 25 millions de voyageurs), suivie de Lille Europe (≈ 7 millions). Les autres gares régionales (CHR, Roubaix, Porte de Douai) présentent un trafic nettement plus faible.

La grande majorité des flux régionaux et interrégionaux convergent vers Lille Flandres, ce qui en fait une gare pivot du nord de la France. Cette information est essentielle pour le dimensionnement des infrastructures locales et l'organisation des correspondances.

Dynamiques locales à Lyon et Marseille

Les graphiques des gares lyonnaises et marseillaises mettent en évidence un schéma similaire :

À Lyon, la Part-Dieu concentre près de 85 % du trafic local, confirmant son statut de gare principale, loin devant Perrache et Vaise.

À Marseille, la gare Saint-Charles (non affichée ici mais connue comme principale) est complétée par Marseille-Blancarde, seule à afficher une fréquentation notable.

Ces observations soulignent la forte polarisation des flux autour d'une seule gare centrale par métropole.

Les autres grandes gares de province

Les gares de Bordeaux Saint-Jean, Nantes et Montpellier Saint-Roch conservent une fréquentation importante (entre 15 et 22 millions de voyageurs). Elles jouent un rôle clé dans la mobilité interrégionale, reliant le nord et le sud-ouest du pays.

Données temps réel (Lille Flandres)

Le graphique "Nombre de départs à Lille Flandres par heure" met en évidence les créneaux de plus forte activité. Même si les données encore récentes ne couvrent que quelques jours, on observe :

- Une activité stable la journée
- Des pics entre 17h et 20h, correspondant aux retours de travail

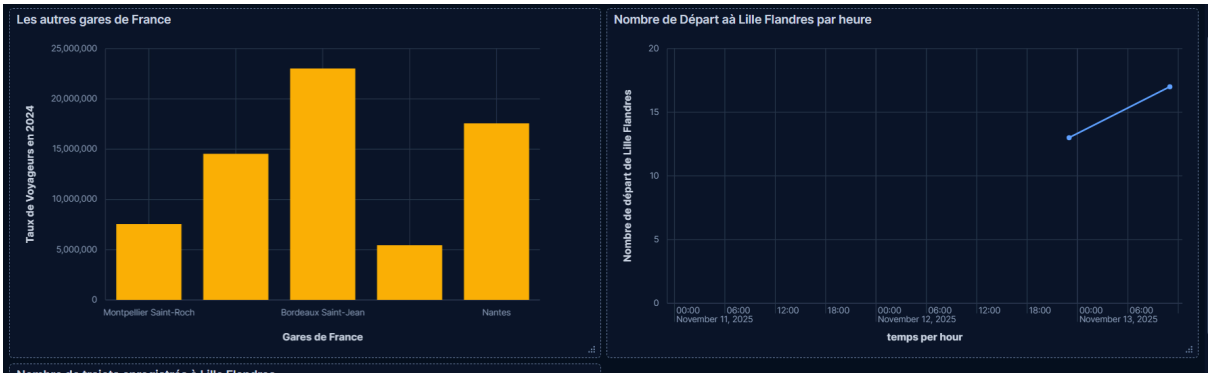
Et un léger décalage horaire dans Kibana dû au stockage UTC les départs apparaissent parfois décalés d'une heure, il faut donc paramétrer Kibana sur le fuseau horaire Europe/Paris pour fiabiliser les analyses temporelles.

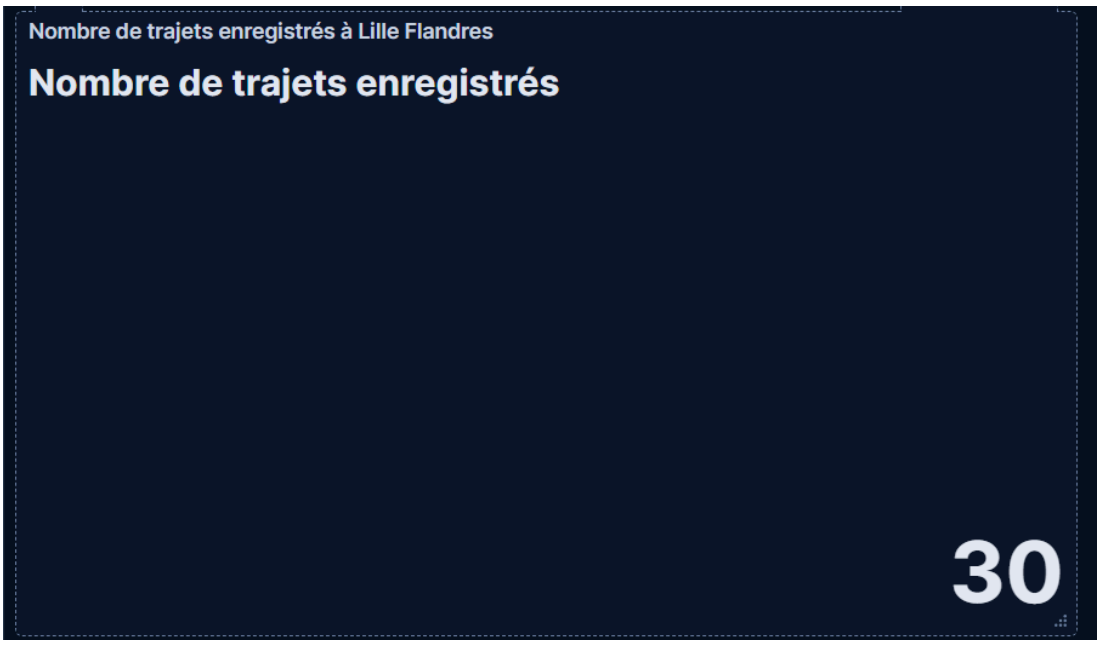
Lecture globale

Ces visualisations permettent d'obtenir une vue unifiée de la mobilité nationale, depuis les grands pôles (Paris, Lyon, Lille, Marseille) jusqu'aux gares régionales.Elles montrent la corrélation entre le poids économique d'une métropole et son volume de voyageurs, ainsi que l'importance des hubs comme Lille Flandres dans le maillage ferroviaire français.


Les 5 gares les plus visitées de France
Taux de Voyageurs en 2024
300,000,000
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
0
Paris Gare du Nord
Paris Gare de Lyon
Paris Est
Gares de France
Les 5 gares les plus visitées de la MEL
Taux de Voyageurs en 2024
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
Lille Flandres
Lille Europe
Lille CHR
Roubaix
Lille Porte de Douai
Gares de la MEL


Les 5 Gares plus plus visitées de Lyon en 2024
Taux de Voyageurs en 2024
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
Lyon Saint-Paul
Lyon Perrache
Lyon Vaise
Lyon Gorge de Loup
Lyon Part Dieu
Gares de Lyon
Les 5 Gares Les plus visitées de Marseille en 2024
Taux de Voyageurs en 2024
20,000,000
15,000,000
10,000,000
5,000,000
0
Marseille Blancarde
Marseille-en-Beauvaisis
Sainte-Marthe en Provence
Gares de Marseille


Les autres gares de France
Taux de Voyageurs en 2024
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
Montpellier Saint-Roch
Bordeaux Saint-Jean
Nantes
Gares de France
Nombre de Départ aà Lille Flandres par heure
Nombre de départ de Lille Flandres
20
15
10
5
0
00:00
November 11, 2025
06:00
12:00
18:00
00:00
November 12, 2025
06:00
12:00
18:00
00:00
November 13, 2025
06:00
temps per hour


Nombre de trajets enregistrés à Lille Flandres
Nombre de trajets enregistrés
30

LIMITES ET PISTE D'AMÉLIORATION

Bien que le projet atteigne ses objectifs principaux, plusieurs limites techniques et fonctionnelles ont été identifiées lors de sa mise en œuvre.

Limites des données

- **Fiabilité de l'API SNCF** : le service ne fournit pas toujours des flux continus et certaines gares peuvent apparaître incomplètes ou avec des métadonnées manquantes (absence de coordonnées ou d'étiquettes).

- **Incohérences ponctuelles** : certains champs de l'API peuvent contenir des formats différents (texte, numérique ou null), nécessitant un nettoyage supplémentaire avant indexation.

Limites techniques

- **Problème de décalage horaire** : une différence a été observée entre les horaires réels et ceux affichés dans Kibana, en raison d'un décalage entre l'heure UTC (stockée dans Elasticsearch) et l'heure locale (Europe/Paris).
  Ce décalage peut donner l'impression que certains trains partent une heure plus tôt ou plus tard selon la période (heure d'été ou d'hiver).
  Une solution consisterait à forcer le fuseau horaire local lors de la génération du champ **timestamp**, ou à ajuster le paramétrage de Kibana pour afficher l'heure selon la zone souhaitée.

- **Performance et scalabilité** : bien que la fréquence d'exécution (toutes les 15 minutes) soit adaptée à ce cas d'usage, une montée en charge avec plusieurs gares nécessiterait un système de traitement asynchrone (par exemple via Kafka ou Logstash).

Pistes d'amélioration

- Étendre la collecte à d'autres gares stratégiques (Lille Europe, Paris-Nord, Lyon-Part-Dieu, etc.) pour comparer les flux de voyageurs ;

- Ajouter une alerte automatique (via webhook ou email) en cas de perturbation importante dans le trafic.

Conclusion

Ce projet a permis de démontrer la mise en œuvre concrète d'un **pipeline de données temps réel** appliqué au domaine de la mobilité ferroviaire. Grâce à la combinaison de **Python**, **Elastic Cloud** et **Kibana**, il a été possible de créer une solution automatisée capable de :

- Collecter des données hétérogènes (CSV, API) ;
- Les nettoyer, enrichir et indexer dans un environnement centralisé ;
- Les visualiser sous forme de **dashboards dynamiques** exploitables pour l'analyse du trafic.

L'étude a également mis en lumière l'importance de la **qualité des données**, du **choix du fuseau horaire** et de la **gestion des erreurs** dans tout processus de data engineering. En dépit de certaines limites techniques, la solution développée constitue une base solide pour de futurs projets d'analyse prédictive ou d'optimisation de la mobilité.À terme, l'intégration de modèles statistiques ou d'apprentissage automatique permettrait d'aller au-delà du suivi descriptif pour anticiper les pics de fréquentation et améliorer la planification du transport ferroviaire.