SIMPLON.CO

KEOLIS

Data Engineer
09/2024 – 03/2026

Rendu E2

Encadrement de la conception d’un projet data : Mise en situation professionnelle

Table des matières

Introduction

Au sein du groupe Keolis, l'usage de la donnée a pris une place grandissante ces dernières années. Face à l'évolution des besoins et des attentes, notamment pour **optimiser l'exploitation** des réseaux de transport en commun, la donnée est devenue un véritable levier de performance au service des filiales.
Dans la Direction Régionale Nord-Est, cette dynamique s'est traduite par une volonté forte du directeur d'exploitation : placer l'analyse des données au cœur des décisions, pour **piloter** les opérations avec plus de précision et d'agilité. C'est dans ce contexte qu'un projet a vu le jour : **concevoir une collecte des données**, pensée pour alimenter des rapports Power BI plus performants et plus pertinents.

Derrière ce projet technique se cache en réalité une architecture plus vaste, **une architecture en médaillon**, construite en plusieurs couches successives qui, ensemble, assurent la solidité et l'efficacité des flux de données. J'en évoquerai brièvement les contours dans ce chapitre, mais pour une compréhension complète, le lecteur pourra se reporter au **glossaire** pour un aperçu des termes essentiels, avant de découvrir les **détails dans la partie E7**.

Enfin, parce qu'un projet data ne se limite pas à la technique, une documentation complète a accompagné chaque étape pour cadrer la gestion, les choix méthodologiques et les livrables attendus.

A Contexte et présentation du projet

Dans le cadre de la gestion des transports en commun, les filiales de la DRNE s'appuient sur les données issues des SAE (Systèmes d'Aide à l'Exploitation) pour piloter et analyser la circulation de leurs véhicules. Ces systèmes fournissent des informations essentielles telles que les horaires, les positions GPS, ou encore les incidents de circulation. Selon les fournisseurs, les modalités de récupération des données varient : certaines proposent une API, d'autres mettent à disposition des fichiers plats accessibles via un serveur.

L'objectif général de ce projet, **spécifique, mesurable, atteignable, réaliste et temporellement défini (SMART)**, est de mettre en place un **flux automatisé** permettant de **récupérer, stocker et exploiter** les données de circulation d'une filiale dans une logique de **gouvernance** et de **valorisation** des données.

Pour illustrer la mise en œuvre d'un projet data en environnement professionnel, j'ai choisi de présenter le flux que j'ai développé pour récupérer les données d'un SAE particulier, utilisé par l'une des filiales de la DRNE. Ce choix permet de détailler l'ensemble de la démarche projet appliquée, de l'expression de besoin jusqu'à la recette technique.

Ce projet s'inscrit dans des **enjeux métiers et stratégiques** de fiabilisation des indicateurs de suivi, d'optimisation des ressources, et d'amélioration du pilotage opérationnel. Il répond également à des objectifs de **traçabilité, sécurisation et pérennisation des flux**.

Les **parties prenantes principales** de ce projet sont :

- Les référents métiers de la filiale utilisatrice,
- Le service data de la DRNE (il n'y a que moi actuellement),
- Le fournisseur du SAE.

L'ensemble des livrables produits – grille d'expression de besoin, documentation technique, procès-verbal de recette – seront détaillés dans les sections suivantes, afin de mettre en évidence la rigueur méthodologique appliquée tout au long de ce projet.

B Analyse du besoin (C1)

1 Contexte et enjeux

Le besoin exprimé par la direction technique est de **mettre à disposition des données sous forme de fichiers plats (CSV)** dans un Blob Storage Azure, selon **des règles de formatage strictes**, permettant leur intégration automatique dans la plateforme centrale de traitement au niveau du groupe.
Ce projet s'inscrit dans une démarche plus large de **valorisation de la donnée**, de **fiabilisation des flux** et de **réduction des tâches manuelles**.
Le traitement s'appuie principalement sur **l'écosystème Microsoft Fabric** et intègre les données brutes provenant du SAE, via un serveur SFTP.

2 Objectifs et périmètre fonctionnel

Objectifs SMART

- **S**pécifique : Fournir un flux automatisé pour la mise à disposition des données dans un Blob Azure au format CSV.
- **M**esurable : Atteindre un taux de conformité de 95 % minimum sur les fichiers transmis (encodage, typage, séparateurs…).
- **A**mbitieux : Réduire le temps de traitement humain à zéro en automatisant l'ensemble du flux.
- **R**éaliste : Projet basé sur des outils maîtrisés (PySpark, Microsoft Fabric, Azure Blob).
- **T**emporellement défini : Mise en production effective janvier 2025.

Périmètre fonctionnel

- **Entrée** : Fichiers CSV quotidiens générés par le fournisseur, récupérés via SFTP.
- **Traitement** : Nettoyage, typage, dédoublonnage et normalisation via Notebook PySpark dans Fabric.
- **Stockage intermédiaire** : Données en Lakehouse (bronze → silver).

- **Sortie** : Fichier CSV structuré, encodé en UTF-8, exporté vers Azure Blob (`input/sae/`), avec nommage normé (`passages_sae_xxxx_yyyymmdd.csv`), séparateurs échappés, retour à la ligne Windows (CRLF).

3 Étude d'opportunités et faisabilité

Opportunités identifiées

- Centralisation des données dans un format exploitable automatiquement.
- Automatisation complète des tâches récurrentes → réduction du risque d'erreur humaine.
- Amélioration de la qualité de données livrées à la plateforme centrale.
- Valorisation interne des données via le Lakehouse.

Contraintes à prendre en compte

- Structure de fichier imposée par le cahier des charges (fort niveau de précision).
- Dépendance aux flux SFTP et à leur fiabilité.
- Nécessité d'avoir une supervision pour détecter les anomalies en production.

Étude de faisabilité (qualité, coûts, délais)

Dimension	Évaluation
Qualité	Contrôles mis en place dans les notebooks, règles de validation automatisées
Coûts	Coût réduit grâce à l'utilisation de composants Fabric existants (Notebook, Pipeline, Dataflow)
Délais	Projet réalisé en 10 à 12 semaines de développement, puis 2 semaines de recette
Moyens	Alternant ingénieur de données + supervision technique du responsable projets IT + infrastructure existante Azure

4 Architecture visée

Le flux de traitement repose sur une **architecture cloud basée sur Microsoft Fabric**, et s'exécute quotidiennement en mode automatisé.

- **Étapes du flux** :
    - Collecte depuis le **SFTP**
    - Traitement dans un **Notebook PySpark** (nettoyage, typage, renommage)
    - Stockage en **tables delta dans un Lakehouse Fabric**
    - Export final en CSV vers **Azure Blob**, avec les spécifications imposées
- Le pipeline inclut des mécanismes :
    - **d'alerte** par mail (succès ou erreur),

- de **validation** automatique du fichier final,
- d'**exécution planifiée** quotidienne.

Un second flux, dérivé du premier, produit une table complémentaire à destination du directeur d'exploitation DRNE pour son usage.

5 Analyse RICE

Critère	Valeur estimée
Reach	50 utilisateurs directs ou indirects concernés
Impact	Fort : meilleure qualité de données, meilleure performance des rapports Power BI et automatisation complète (3/3)
Confidence	Élevée : infrastructure maîtrisée, tests validés (>95 %)
Effort	Moyen : 25 jours.homme cumulés

Score RICE = (50 × 3 × 0.85) / 25 = 5.1

Le projet présente un **excellent rapport valeur/effort**, avec une automatisation efficace, une intégration fluide dans l'écosystème Microsoft, et une réponse claire au besoin métier initial.

6 Cadrage fonctionnel détaillé

Élément	Contenu
Source	CSV via SFTP sécurisé
Données sensibles	Seule donnée sensible : matricule conducteur (Registre RGPD à respecter) présente uniquement dans les données quotidiennes.
Rétention	Bronze : 1 an ; Silver : 1 an (données journalières), 2 ans (données mensualisées), 5 ans (données annualisées)
Fréquence / volumétrie	1 fichier csv par jour ; pas d'historique à récupérer
Traitements attendus	Dédoublonnage, typage, nettoyage, renommage
Sortie attendue	CSV nommé `passages_sae_xxxx_yyyymmdd.csv`, encodé UTF-8, CRLF
Accès utilisateurs	Lecture seule pour les utilisateurs métier

7 Accessibilité et évolutivité

Côté technique :

- **Cloud Fabric** : accessible à distance.
- **Notebooks PySpark** : souplesse, lisibilité.
- **Documentation intégrée** : facilitant la reprise du projet.

Côté métier :

- **Nomenclature explicite** : fichiers et colonnes.
- **Dictionnaire de données** : prévu pour accompagner l'usager.
- **Version DRNE** : adaptée à un besoin spécifique.

8 Hypothèses et perspectives

Les flux quotidiens resteront stables à court terme.

- Le système pourra être généralisé à d'autres SAE.
- Aucun besoin d'interface spécifique : l'accès aux fichiers suffit.
- Les outils Fabric sont adaptés à une montée en charge.

En conclusion, le projet est **techniquement réaliste**, **évolutif**, **robuste** et **aligné sur les besoins métiers**. Il répond aux enjeux de qualité, de conformité RGPD, et de valorisation des données, tout en restant **accessible** aux utilisateurs finaux et **maintenable** par les équipes techniques.

Il offre également une **base réplicable** pour d'autres flux similaires dans le groupe, ouvrant la voie à une stratégie d'industrialisation de la donnée à plus grande échelle.

C Cartographie des données (C2)

Les données du projets sont toutes structurées. Elles sont présentes au format **csv** dans la **source**, au format csv dans la couche bronze dans le **lakehouse**, au format **parquet** dans le même lakehouse dans la couche silver et au format csv dans la couche gold sur le **blob** du compte de stockage Azure. Le tableau suivant donne un extrait des données de chaque couche :

Source	**Couche Bronze**	**Couche Silver**		**Couche Gold Impulse**	
CSV	CSV	TABLE DELTA		CSV	
		COLUMN_NAME	**DATA_TYPE**		
Date	Date	Date	varchar	date_exploitation	
Jour	Jour	Jour	varchar		
Periode	Periode	Periode	varchar		
Ligne	Ligne	Ligne	varchar		
SV	SV	SV	varchar		
NumeroCourse	NumeroCourse	NumeroCourse	int	actual_trip_id	trip_id
Vehicule	Vehicule	Vehicule	int	vehicle_id	
Arret	Arret	Arret	varchar		
CodeArret	CodeArret	CodeArret	varchar		
IndicePassageArret	IndicePassageArret	IndicePassageArret	int	actual_stop_sequence	
Sens	Sens	Sens	int	direction_id	

On a également un extrait du glossaire général des données :

Nom champ	Description	Exemple
Date	Date du relevé	29/01/2025
Jour	Nom du jour du re	mercredi
Periode	Période de relevé de la donnée	Septembre 2024 - PS (VV01)
Ligne	Ligne de bus	L10
SV	Numéro du service	333
NumeroCourse	Numéro de la course	444
Vehicule	Numéro du véhicule	555
Arret	Nom de l'arrêt	Chapelle
CodeArret	Code de l'arrêt	CHL01
IndicePassageArret	Indice de l'arrêt dans le trajet	1
Sens	Sens de passage	1

Les accès aux données sont programmés de la manière suivante :

Rôle utilisateur	Accès donné	Niveau (Lecture/Écriture)
Exploitation DRNE	Lakehouse tables	Lecture
Plateforme centrale Groupe	Blob Azure	Lecture + Intégration
Équipe IT	Fichiers CSV, Lakehouse tables	Lecture/Écriture

D Spécifications techniques et architecture cible (C3)

1. Outils et technologies utilisés

Le projet repose sur une architecture cloud intégralement hébergée sur Microsoft Fabric, associée aux outils suivants :

Outil / Technologie	Rôle dans le projet
Microsoft Fabric (Lakehouse)	Stockage des données brutes et transformées (bronze, silver)
Notebook PySpark	Traitement des données : nettoyage, typage, dédoublonnage
Azure Blob Storage	Stockage final des fichiers CSV normés (gold)
Pipeline Fabric	Orchestration des flux : exécution, planification, alertes
SFTP sécurisé	Source des fichiers quotidiens

Ces outils sont maîtrisés en interne et adaptés à l'industrialisation des flux.

2. Architecture mise en place

L'architecture effective est à l'image de celle visée. Je joins en annexe les différentes représentations de celle-ci.

3. Contraintes techniques

Les principales contraintes identifiées sont :

- Respect strict de la **structure des fichiers** exigée par le cahier des charges (nomenclature, séparateurs, encodage, typage des colonnes).
- **Fiabilité** des flux SFTP, avec nécessité d'une supervision active.
- **Pérennité** du stockage et des traitements, intégrés dans l'écosystème Fabric.
- Aucun traitement RGPD particulier hors respect du matricule conducteur. Les traitements de données RGPD est détaillé dans le rendu E7.
- Eco-responsabilité : Microsoft Fabric est un environnement mutualisé qui permet **d'optimiser** les ressources cloud. De plus les copies de données sont limitées puisque la couche silver sert à la fois à alimenter le demandeur initial et le service exploitation pour ses propres analyses. Enfin les traitements sont planifiés la nuit afin de répartir la charge par rapports aux flux qui ont des contraintes horaires plus exigeantes.
- **Accessibilité** des livrables : la documentation produite (procédures, dictionnaire de données, schémas) est mise à disposition en format PDF accessible, compatible avec les outils d'assistance numérique.

E Veille technologique et réglementaire (C4)

L'entreprise évoluant dans un environnement entièrement basé sur Microsoft, et les outils data étant désormais réunis au sein d'une interface unifiée avec Microsoft Fabric, j'ai choisi d'orienter ma veille sur l'**outil principal** que j'utilise au quotidien. Ce choix se révèle d'autant plus pertinent que les fonctionnalités de ces outils évoluent à un rythme soutenu, avec des mises à jour quasi quotidiennes.

🗓️ **Cadence hebdo : vendredi après-midi, 2 h**
📌 **Thématique : nouveautés Data Fabric dans le cadre de notre utilisation de Microsoft.**

🔍 **Principales évolutions** :

- **Oct–Nov '24** : export/import pipelines, Copilot, monitoring EventHouse.
- **Fév '25** : Dataflow Gen2 devient paramétrable, « Enregistrer sous », migr. Gen1 → Gen2.
- **Fév '25** : support ZIP compression dans Copy Activity.
- **Mars '25** : Dataflow Gen 2 intégré dans le versionning Git.
- **Fév–Avr '25** : monitoring mirroring, journaux audit.
- **Avr-Mai '25 :** Copy Data simplifie la collecte de données on-premises.
- **Juin '25 :** Real-Time Intlligence amélioré.
- **Sept '25 :** apparition de Copilot dans Dataflow Gen2.
- **Oct '25 :** intégration de Key Vault dans Fabric.

🧭 **Sources** :

- Microsoft Learn release plan
- Next Decision blogs
- Azure blog officiel
- Interactions sur Reddit technique

F Suivi opérationnel et pilotage du projet (C6)

Le suivi du projet s'est organisé autour de revues hebdomadaires avec le chef de projet SI et le DSI. Selon l'avancement, le directeur d'exploitation, principal utilisateur, participait pour valider les résultats intermédiaires et ajuster les priorités.

J'assurais la préparation de ces réunions via un ordre du jour partagé dans ClickUp, précisant les sujets, priorités et points en attente. L'équipe pouvait y ajouter ses remarques en amont, favorisant ainsi la participation collective. Le compte rendu, mis à jour en direct, servait de fil conducteur et garantissait la traçabilité des actions d'une semaine à l'autre.

Bien que la communication avec les prestataires soit coordonnée par le chef de projet SI, je contribuais à l'alignement technique en fournissant des spécifications claires et des choix d'outils documentés.

Le suivi budgétaire, centré sur les ressources cloud (stockage et calcul), s'appuyait sur des alertes automatiques et un reporting régulier au DSI.
Grâce à ClickUp, l'avancement global restait visible à tout moment via des vues Gantt ou liste, assurant une gestion fluide et transparente.

Conclusion

Les méthodes de communication et les outils collaboratifs de l'équipe SI de la DRNE de Keolis ont permis de piloter efficacement ce projet, représentatif des pratiques déployées durant mon alternance. J'y ai transposé les principes de gestion issus de mon expérience en production industrielle : structuration du travail, rituels de suivi et contrôle qualité appliqué aux flux de données.
Teams, OneDrive et ClickUp ont assuré la traçabilité et la coordination nécessaires à une conduite de projet fluide et réactive. Ce premier flux a servi de projet pilote pour valider l'architecture cible : intégration entre Microsoft Fabric, Power BI et Lakehouse, avec des performances de chargement optimisées.

En somme, ce projet a posé les bases d'une production de données fiable, évolutive et maîtrisée — ma première véritable « collection » data.

Glossaire

API : application programming interface ou « interface de programmation d'application » est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités.

Architecture en médaillon : c’est une méthode d’organisation largement utilisée dans le monde de la data, car elle permet de structurer les données en différentes couches, en évitant les doublons et en garantissant la fiabilité des informations. Le principe est simple : on commence par collecter et stocker les données brutes dans une **couche bronze**. Une première transformation les nettoie et les enrichit pour alimenter la **couche silver**. Enfin, une dernière étape prépare la **couche gold**, qui met à disposition des utilisateurs finaux des données prêtes à l’emploi, au format adapté à leurs besoins métier.

DRNE : Direction Régionale Nord Est, elle regroupe toutes les filiales et activités du nord est de la France, de Dunkerque à Strasbourg en passant par l’Aisne.

La DSI : la direction des systèmes d’informations

Le DSI : le directeur des systèmes d’informations

RGPD : Règlement général de protection des données.

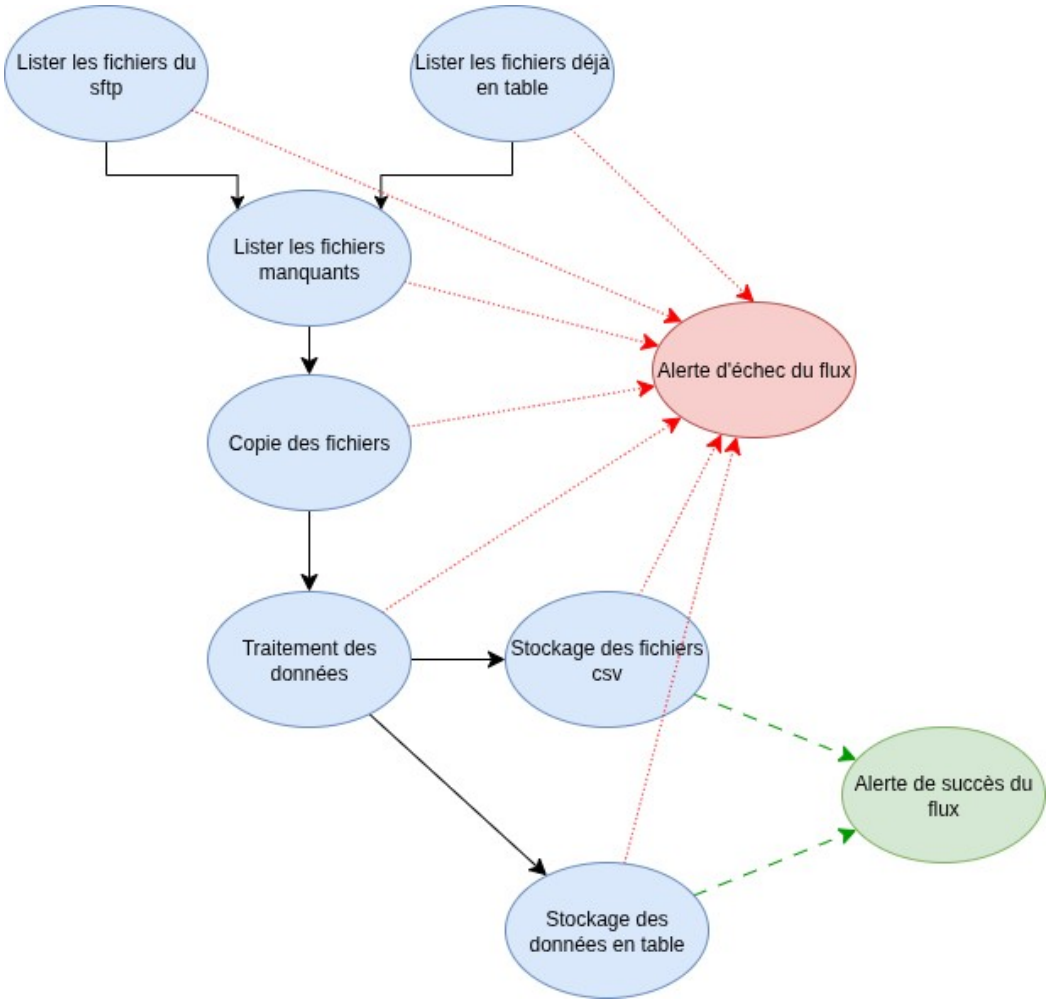
SAE : Système d’Aide à l’Exploitation.

SFTP : Secure File Transfer Protocol est un protocole de transfert de fichiers qui exploite un ensemble d'utilitaires qui fournissent un accès sécurisé à un ordinateur distant par communication sécurisée. Il est considéré comme la méthode optimale pour le transfert de fichiers sécurisés.

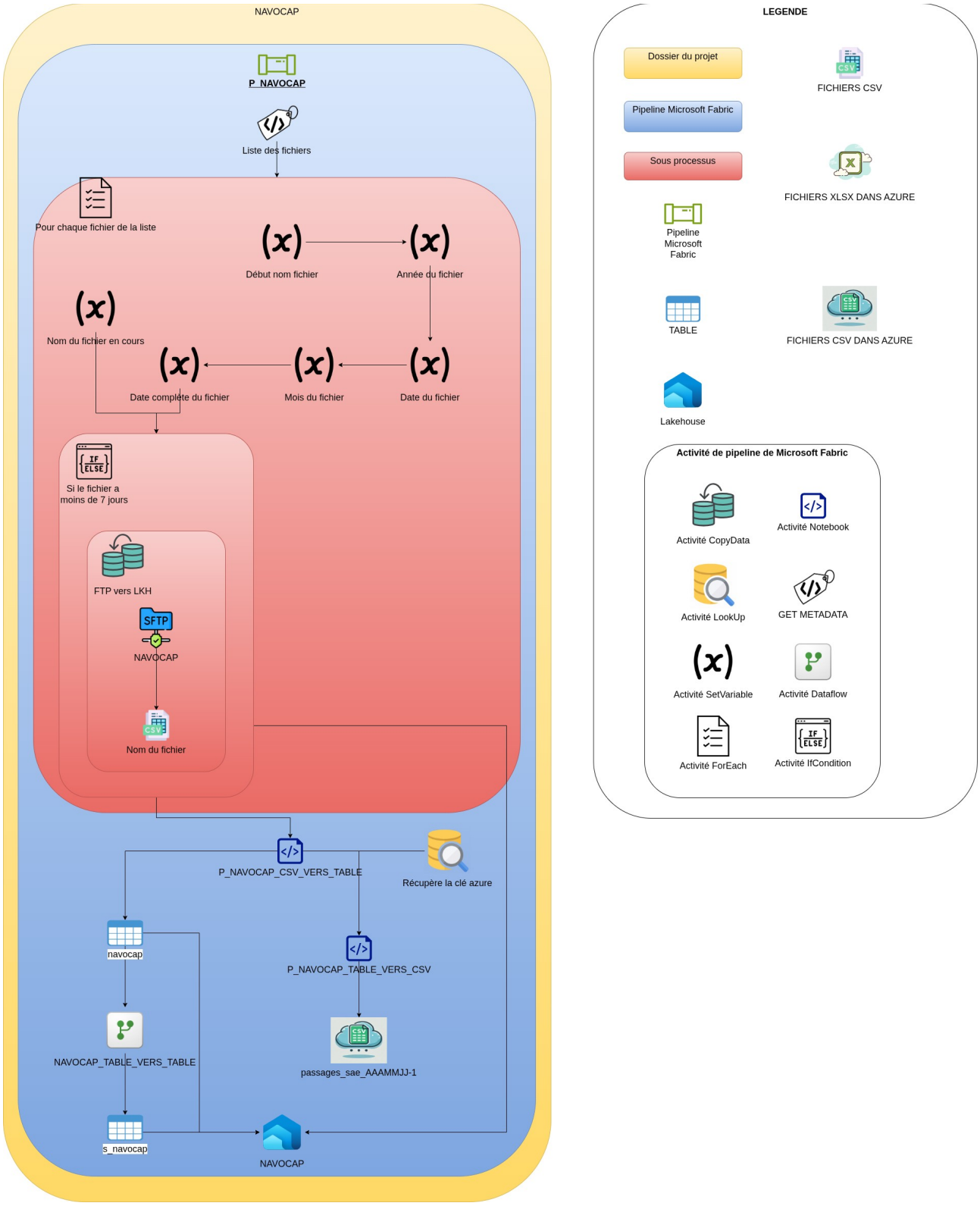
Annexes

Architecture mise en place

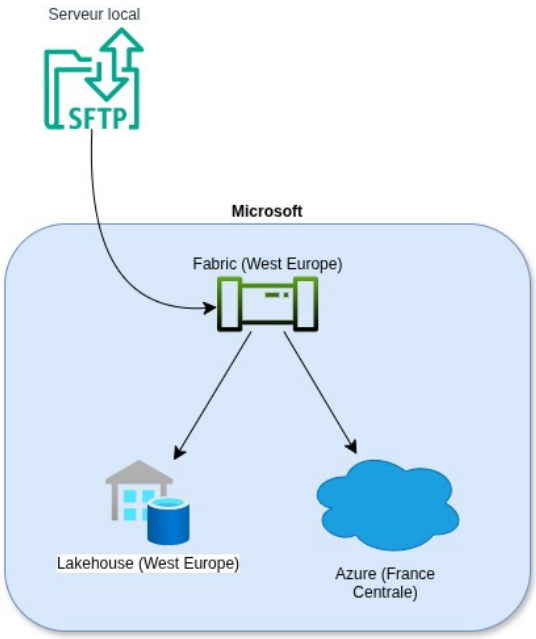
- La représentation fonctionnelle



- La représentation applicative



- La représentation d'infrastructure



- La représentation opérationnelle