Atelier – Architecture Décisionnel (DataMart)

Rakib SHEIKH – TRDE704
noobzik@pm.me – EPSI Paris & Arras
Cours pour :

- I1 EISI
- I1 ECDPIA
- I1 ESI CYBER

L'objectif

- 14h de cours ensemble
- Vous faire pratiquer au déploiement d'une achitecture Décisionnel
- Cours Magistral à 10% / 90 % Pratique (prévisionnel)

- Prérequis :
  - Outil ETL (KNIME, Talend Open Studio, Dataiku, Snowflake...)
  - Licence Tableau Software Desktop (Normalement déjà installée)
  - PC avec 16 Go de RAM minimum
  - Au moins 10Go d'espace disque requis
  - Docker

Evaluation

- Module impliqué dans le cadre du MSPR
  - Déploiement d'une architecture ou solution de sécurité / cybersécurité d'une architecture (ESI CYBER) TRPE 833
  - Big Data et Analyse de données TRPE813 (EISI I1)

- Module soumis à un rendu de projet :
  - Rapport + Code Source + Tableau de bord + Lien Github/Gitlab/Bitbucket noobzik@pm.me
  - Sujet : [EPSI Paris + Classe + Groupe] Rendu ATL Datamart – Noms de familles
- Modalités des groupes
  - G1 : Groupe de 3 + 1 groupe de 4 (non négociable)
  - G2 : Groupe de 4 (A vérifier)
  - ECDPIA : ?
  - ESI CYBER : Groupe de 3 (non négociable)
  - Constitution autonome

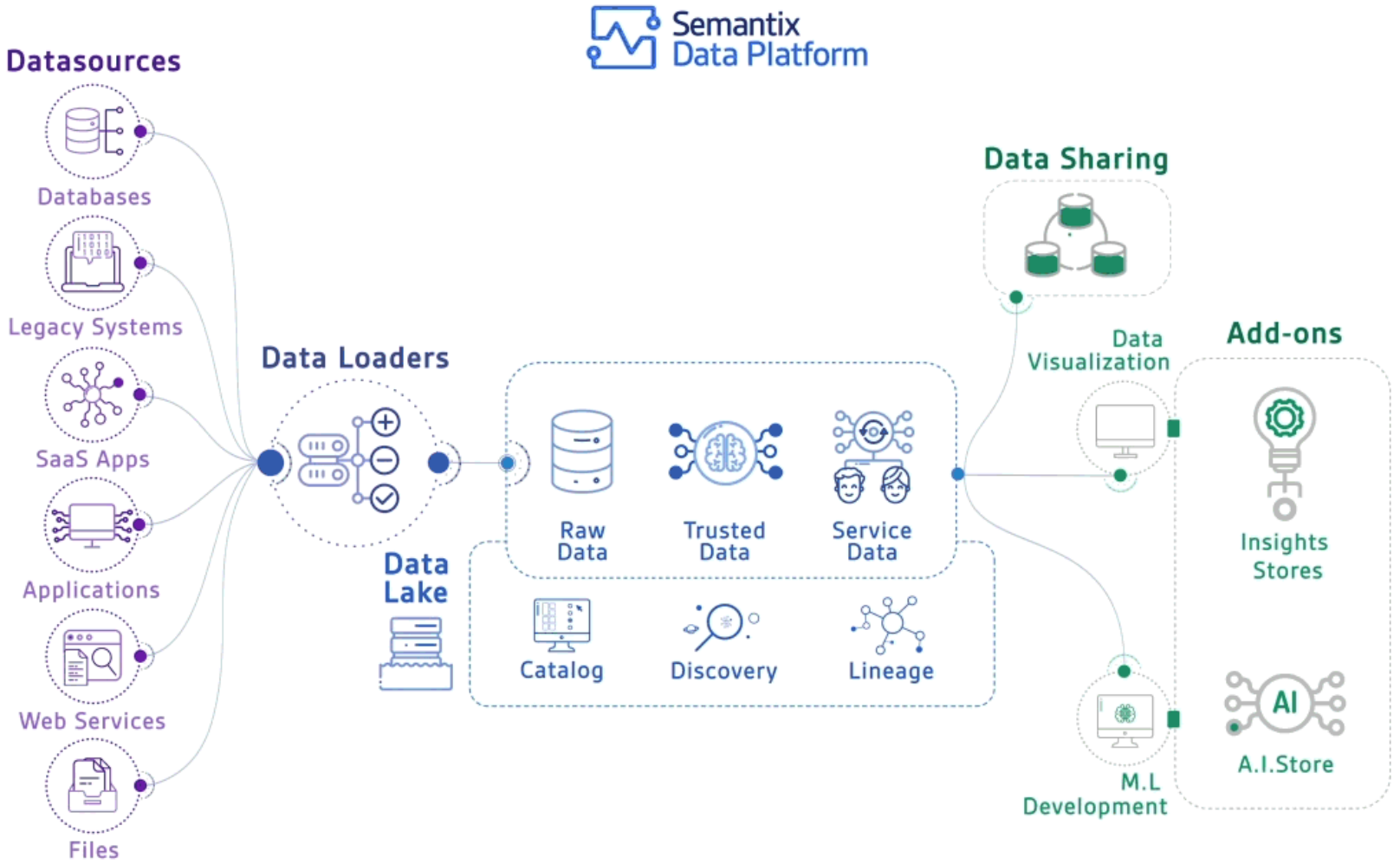
Plan du cours

- Rappels
- TP 1
- TP 2
- TP 3
- TP 4
- TP 5

- Ensemble des TP donne lieu à un rendu d'un code source, tableau de bord et de rapport expliquant votre démarche

Rappels

Chaine de restitution d'un projet
d'informatique décisionnel

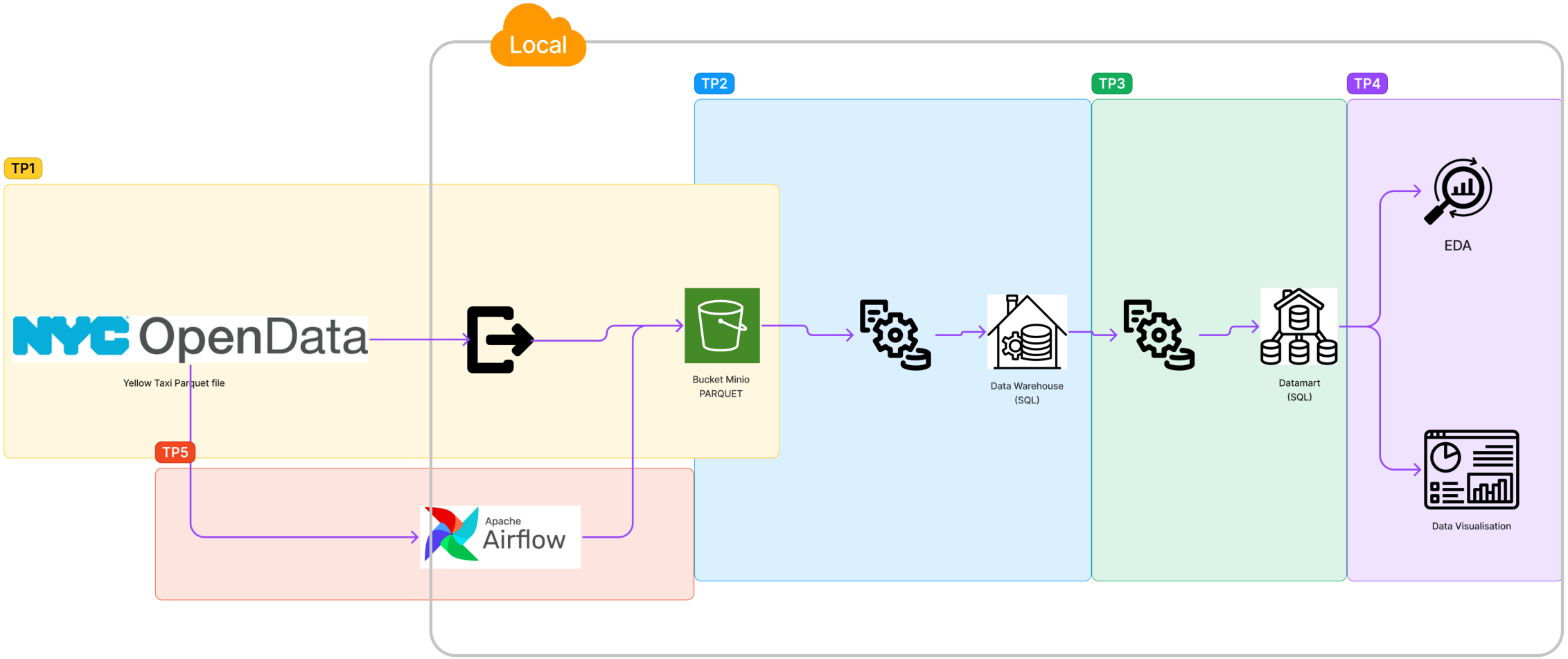

Semantix
Data Platform
Datasources
Databases
Legacy Systems
SaaS Apps
Applications
Web Services
Files
Data Loaders
Data
Lake
Raw
Data
Trusted
Data
Service
Data
Catalog
Discovery
Lineage
Data Sharing
Data
Visualization
Add-ons
Insights
Stores
AI
A.I.Store
M.L
Development

Projet : New York

- L'objectif de ce projet est d'aider une entreprise de VTC basée à New-York à utiliser une partie de ses données pour générer une connaissance.

- Le projet à mettre en œuvre doit répondre au cahier des charges suivants:
  - Automatiser la récupération des données du site de l'Etat de NY vers un Data Lake
  - Appliquer un processus « ETL » pour avoir des données propres et les stocker dans un Data Warehouse.
  - Appliquer un deuxième processus ETL pour avoir des données utilisables en visualisation de dashbord puis stocker dans un Data Mart.
  - Utiliser un outil de Dataviz pour concevoir un dashboard connecté au Data Mart

Architecture du TP à réaliser



TP 1

Business Intelligence

- Nous nous contenterons d'utiliser les données à partir de l'année Janvier 2018

- Nous souhaitons automatiser la création et l'alimentation du Datalake grâce à un export mensuel.
- Les données et les référentiels sont accessibles ici : https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

- Pour cela, nous utiliserons un outil :(Minio, un outil OpenSource dont nous l'utiliserons en tant de Data Lake.)

- Ecrivez deux programme (Python ou Scala) dans le dossier src/data:
  - Le premier récupère tous les datasets de janvier 2023 à aout 2023 (ou plus si dispo)
  - Le deuxième récupère le dernier mois
  - Les données seront téléchargées et stockées dans minio

TP 2

Le système d'alimentation (ETL)

Au choix

Nous allons stocker nos données brutes dans une base de données relationnel SQL sans transformation de données.

- Avec un outil d'ETL :
  - Récupérez les données stockés Minio.
  - Chargez le résultat vers votre base de données.
- Avec du code en Python ou Scala
  - Récupérez les données stockés Minio
  - Chargez le résultat vers votre base de données.

- Le choix de la base de données est libre.
  - Recommandation : PostgreSQL

TP 3

Stockage et DataMarts

Maintenant que vous avez des données qui sont stockés dans une base de données. Il faudra les retravailler pour ajouter une notion de partitionnement afin d'accroitre la performance des requêtes.

- Concevez un modèle en flocon pour les données finaux qui servira au Dashboard.
- Stockez les résultats de votre travail dans une DBMS OLAP dédié au Datamart.
    - Recommandation : PostgresSQL

TP 4

Visualisation de données

1. Avec Tableau ou Power BI:
    - Connectez-vous à votre DBMS Datamart.
    - Faites une EDA (première partie de la visualisation, Notebook Jupyter conseillé)

1. A la suite de votre EDA, vous devriez avoir plus de connaissance sur votre BDD, vous êtes normalement capable identifier les différentes KPI que vous souhaitez restituer.
    - Avec votre outil de BI, concevez vos visualisations de données.
    - Fusionner l'ensemble de vos visualisations de données vers un Dashboard. pour cela, vous allez concevoir votre dashboard.
    - Assurez-vous que votre dashboard se mette à jour automatiquement.

TP 5

Introduction à l'automatisation des tâches

- Nous avons vu dans le cours de Data Science et ML (TRDE706 Paris) que les Data Engineers sont des fainéants.
- Mais pour être fainéant, il faut connaitre les outils d'automatisation.
- Rédiger une DAG dans Apache Airflow en reprenant le code source du premier TP sur la fonction « Récupérer le mois dernier » qui va s'exécuter tous les 1ers du mois