Master 1 – Data Engineering – S2 – 2024/2025

Projet

DAP: Data Analytics Pipeline

Présentation générale

Le projet vise à concevoir et implémenter une plateforme **DAP (Data Analytics Pipeline)** permettant la mise en place d'un pipeline de traitement de données de bout en bout (end-to-end). Ce pipeline doit permettre de :

- Collecter et ingérer des données structurées à partir d'une source (CSV, API ou base SQL),
- Les transformer et les nettoyer progressivement selon le modèle **Bronze → Silver → Gold**,
- Les charger dans une base analytique (locale ou cloud),
- Et de les visualiser via un tableau de bord interactif.

Vous pouvez choisir d'implémenter ce pipeline :

- Soit en **local**, en s'appuyant sur des outils open-source,
- Soit en **cloud**, principalement à travers les services **Azure**.

Le projet peut inclure (sans s'y limiter) les fonctionnalités suivantes :

- **Ingestion automatisée des données** à partir d'une ou plusieurs sources (API, fichiers, SQL, etc.),
- **Stockage multi-niveaux** (Bronze, Silver, Gold) en fichiers ou base de données,
- **Nettoyage et transformation des données** via un moteur de traitement (Pandas ou PySpark),
- **Chargement des données finales** dans une base SQL ou analytique,
- **Création d'un tableau de bord** interactif (Power BI, Streamlit, Tableau Public...),
- **Sécurisation des accès et configuration** (fichiers .env, Azure Key Vault...),
- **Déclenchement automatique** du pipeline via un orchestrateur (ex: Airflow ou Azure Data Factory),
- **Documentation** du pipeline (diagramme de flux, README, etc.).

Spécifications

Les étapes à suivre sont les suivantes :

- **Définir le périmètre fonctionnel du pipeline** (choix du dataset, fréquence, outils, etc.)
- **Proposer une architecture claire** du pipeline (composants, interactions, stockage),
- **Choisir les outils technologiques adaptés** selon les contraintes de votre projet,
- **Développer les scripts d'ingestion, de transformation et de chargement** (ETL),

- **Stocker les données transformées** selon la logique Bronze / Silver / Gold,
- **Créer une visualisation interactive et dynamique** des données finales,
- **Tester** chaque étape du pipeline (unitaires et bout-en-bout),
- **Documenter** le travail réalisé dans un rapport clair.

Technologies

Les étudiants sont libres de choisir les outils adaptés à leur contexte, en les justifiant :

Domaine	**En local**	**Sur Azure Cloud**
Ingestion	Python (requests, pandas)	Azure Data Factory
Traitement	Pandas ou PySpark local	Azure Databricks
Stockage	CSV / Parquet / SQLite / PostgreSQL	Azure Data Lake Gen2
Chargement final	SQL / NoSQl (Mongodb, Cassandra)	Azure SQL / Synapse
BI / Visualisation	Power BI Desktop, Streamlit, Dash	Power BI (service)
Orchestration	Apache Airflow	ADF, Logic Apps
Sécurité	.env + Git	Azure Key Vault, Azure AD

- **Travail en binômes** (max. 2 étudiants)
- **Livrables** :
    - Code du projet (GitHub ou archive),
    - Rapport synthétique (4–8 pages) avec architecture, outils, tests,
    - Présentation orale en fin de semestre (20 min max, démo incluse).

Exemples de projets :

1. Data Pipeline pour l'Analyse des Arnaques en Ligne

Objectif : Collecter, nettoyer, et analyser des données sur les escroqueries signalées (SMS, emails, achats en ligne) pour visualiser les tendances par type, région, et période.
Sources possibles : data.gouv.fr, Kaggle
Innovations :

- Classification automatique des arnaques (via NLP simple)
- Détection d'anomalies dans les zones géographiques

2. Observatoire des Changements Climatiques

Objectif : Mettre en place un pipeline qui suit l'évolution du climat (températures, précipitations, incendies…) dans différentes régions du monde.

Sources : API Météo
Innovations :

- Visualisation interactive des anomalies climatiques
- Analyse prédictive de tendances saisonnières

3. Analyse de Sentiments des Restaurants Locaux

Objectif : Ingestion de commentaires Google/TripAdvisor et classification des restaurants selon les avis clients.
Innovations :

- Scoring des établissements par catégorie (propreté, goût, service…)
- Visualisation des quartiers à fort potentiel gastronomique

4. Pipeline de Prévision de la Circulation Urbaine

Objectif : Traiter des données de trafic en temps réel pour prévoir les congestions et incidents routiers.
Sources : OpenTraffic, HERE API, City data
Innovations :

- Modèle prédictif de congestion
- Affichage dynamique des itinéraires alternatifs

5. Tableau de Bord Éthique de l'IA

Objectif : Collecter des incidents liés à l'usage biaisé de l'IA (recrutement, justice, sécurité) et en dégager des tendances.
Sources : News, GitHub Issues, AI Incident Database
Innovations :

- Analyse textuelle des incidents
- Recommandations pour la régulation éthique