Membres du groupe :

- Maghouti Aymane
- Outmani Ossama
- Elgharbaoui Abdelghafor
- Motassim Hamza

1. Identification des sources :

- Bases de données de brevets publiques :

- Google Patents : Plateforme en ligne fournissant un accès gratuit aux brevets délivrés dans le monde entier. (Offre une API publique avec des limitations)

- USPTO (United States Patent and Trademark Office) : Base de données officielle des brevets américains. (Offre une API publique avec des limitations)

- EPO (European Patent Office) : Base de données officielle des brevets européens. (Offre une API publique avec des limitations)

2. Proposition d'une architecture Big Data basée sur Spark pour l'analyse de données :

- Ingestion des données :

- Apache Kafka : Plateforme de streaming distribuée pour l'ingestion des données en temps réel.

- Stockage :

- Hadoop HDFS (Hadoop Distributed File System) : Système de fichiers distribué pour le stockage de données volumineuses.

- Apache HBase : Base de données NoSQL distribuée pour le stockage de données structurées.

- Traitement et analyse :

- Apache Spark : Moteur de traitement de données rapide et extensible pour l'analyse en mémoire et sur disque.

- Microsoft Power BI : Outils de visualisation de données.

- Justification des choix technologiques :

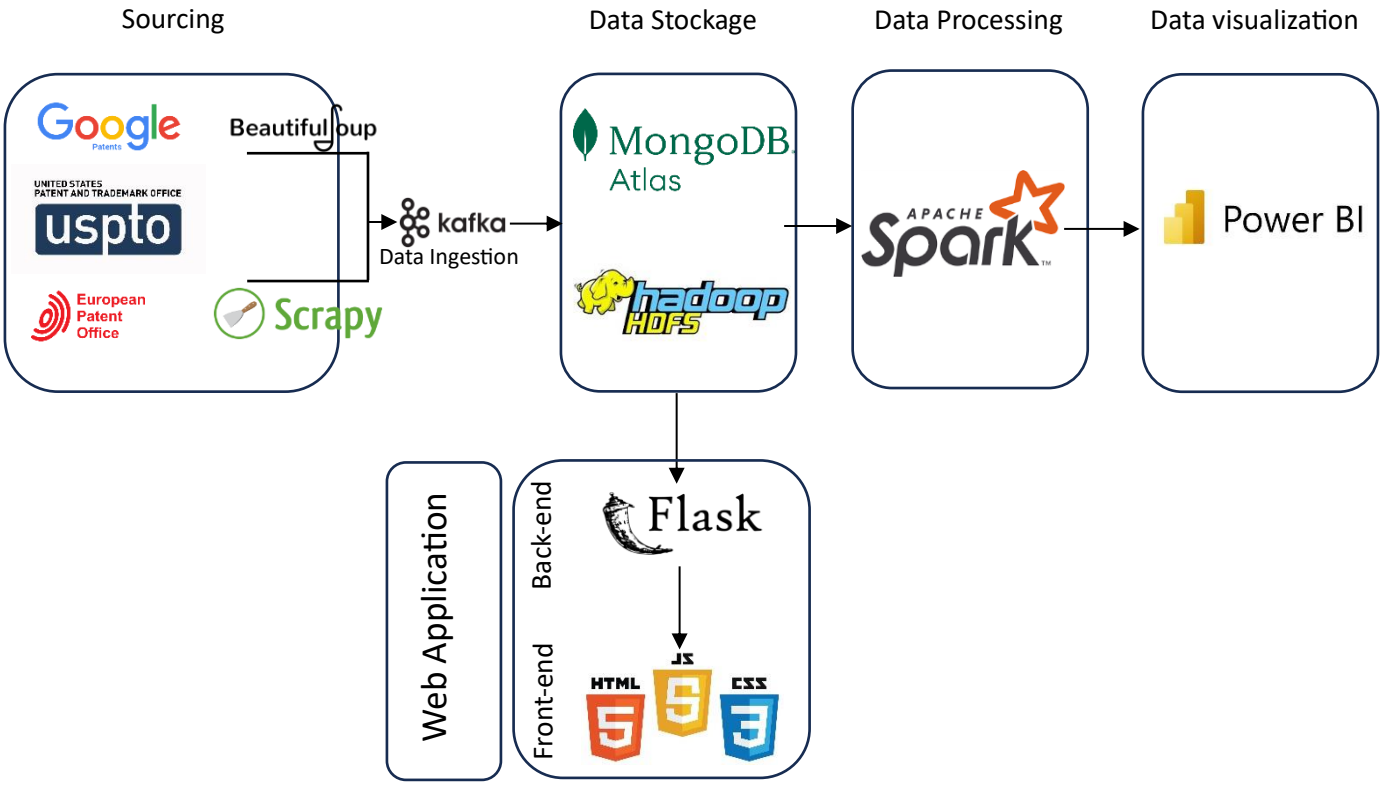
- Apache Kafka permet une ingestion efficace des flux de données en temps réel, ce qui est crucial pour capturer les mises à jour des bases de données de brevets.

- Hadoop HDFS offre une capacité de stockage évolutive pour gérer le volume important de données de brevets.

- Apache HBase permet un accès rapide aux données structurées, ce qui est essentiel pour les requêtes analytiques.

- Apache Spark offre des performances élevées pour le traitement des données distribuées, ce qui accélère l'analyse des brevets.

Architecture Big data :



3. Groupe : Workflow :

- Sourcing :
  - Extraction des données à partir des sources identifiées.
- Stockage :
  - Stockage des données brutes dans Hadoop HDFS.
- Analyse et traitement :
  - Prétraitement des données : nettoyage, normalisation, etc.
  - Analyse des brevets à l'aide d'Apache Spark.
- Visualisation :
  - Création de visualisations interactives pour présenter les résultats de l'analyse en utilisant PowerBI.

4. Retro-planning :

- Identification des tâches :

  - Recherche et sélection des sources de données.

  - Configuration de l'environnement Big Data (installation des technologies, etc.).

  - Développement des scripts d'ingestion des données.

  - Mise en œuvre des pipelines de traitement des données.

  - Développement des visualisations.

- Délais et ressources :

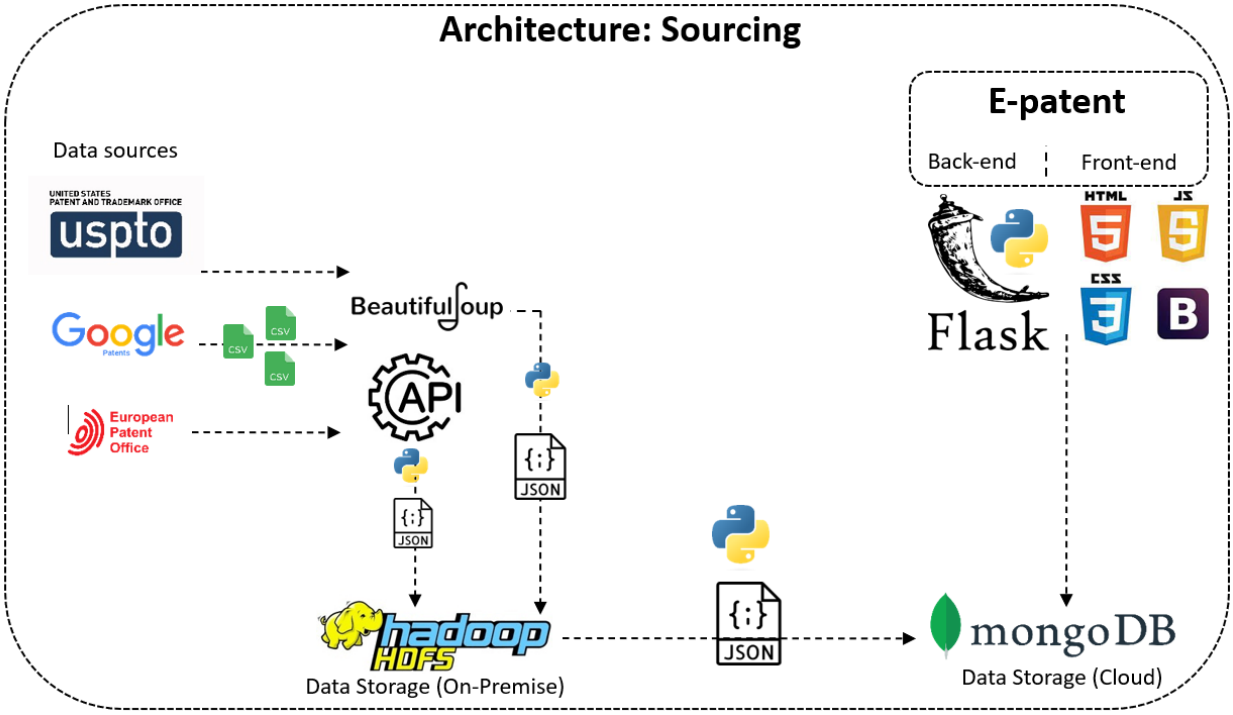
  - Chaque tâche sera affectée à un ou plusieurs membres de l'équipe en fonction de leurs compétences et de leur charge de travail.

  - Les délais seront fixés en fonction de la complexité de chaque tâche et des ressources disponibles.

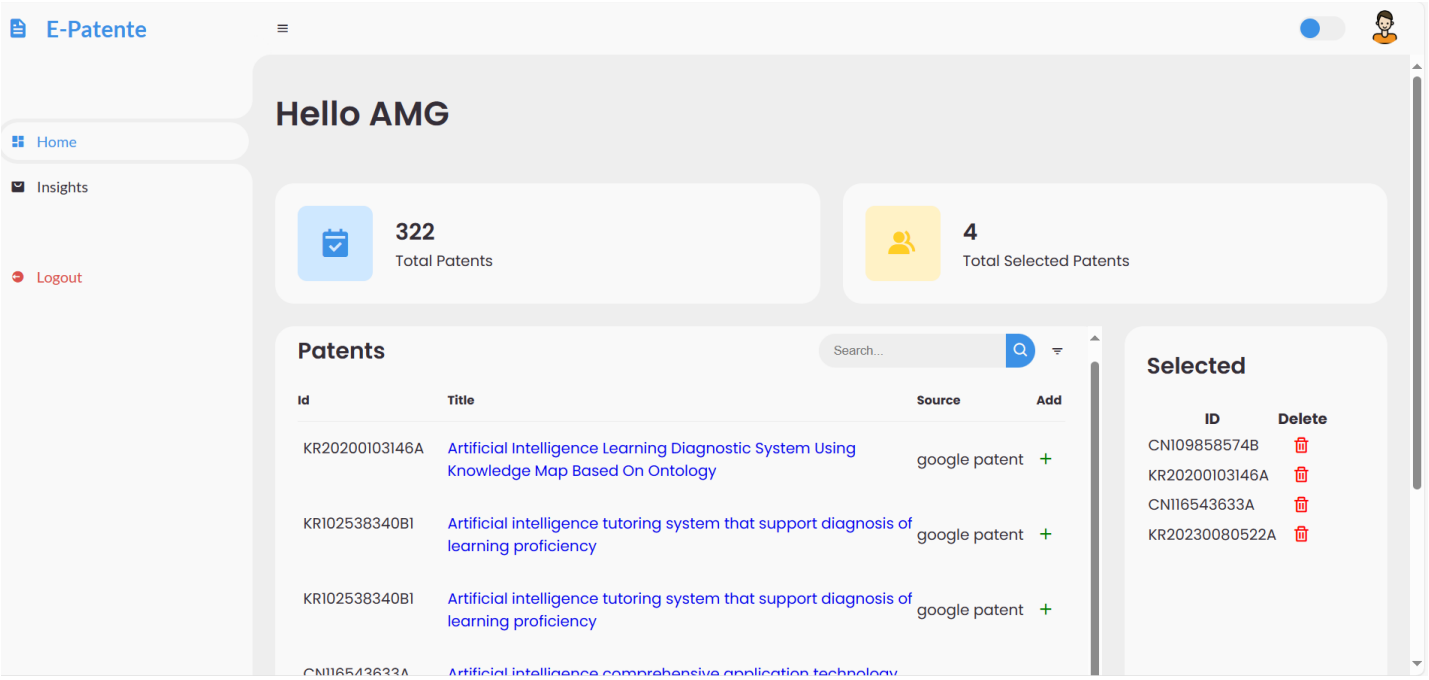
- Suivi et ajustement :

  - Des réunions régulières seront organisées pour suivre la progression du projet et ajuster le rétro-planning si nécessaire en fonction des imprévus ou des retards.

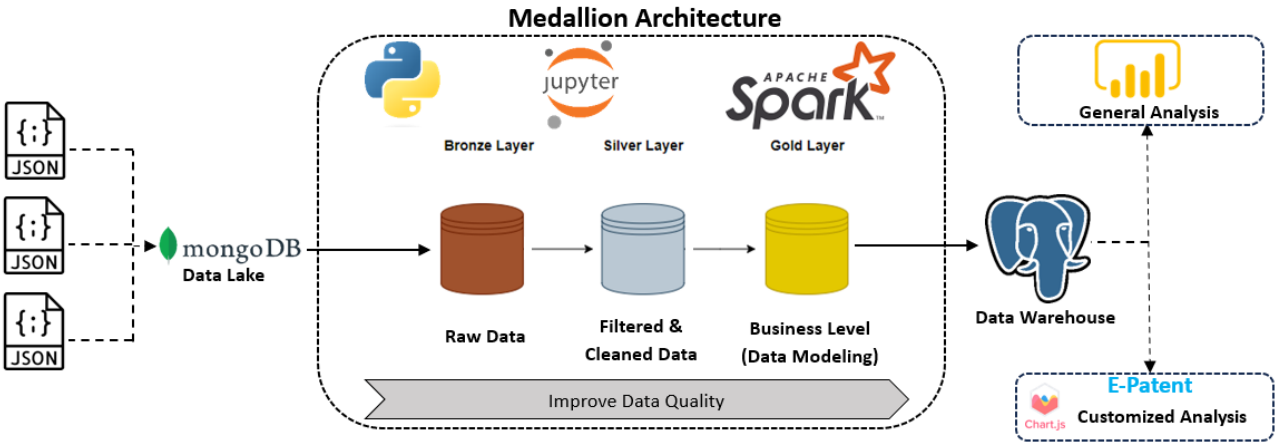
Part 1: Sourcing



The UI:



Partie 2 : Construction du DW avec une architecture en médaillon



- **Le schème en étoile**