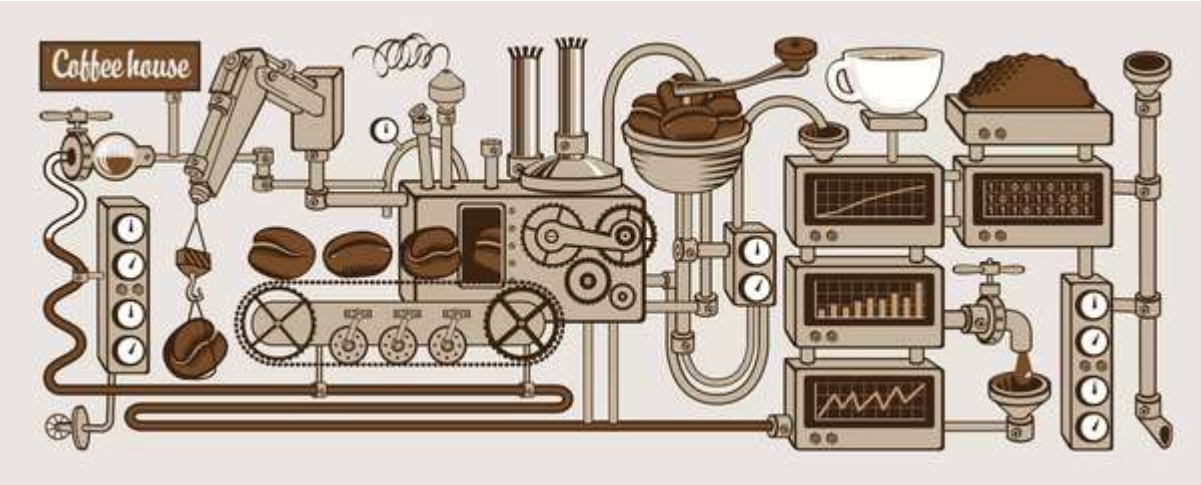
SAYF BEJAOUI
CONSULTANT DATA ENGINNER-SCIENTIST EN ENTREPRISE
FORMATEUR BIG DATA, IA & CLOUD

Projet Pro final
NoSQL et Big Data

Mise en situation

Vous êtes Développeur Data en alternance fraîchement embauché depuis une semaine. Vous avez fait connaissance avec vos collègues, votre nouveau bureau, mais surtout, la machine à café high-tech :



Rien que ça !

Vos missions :

Il est temps de mettre les mains dans le cambouis ! Votre **manager/tuteur** souhaite que vous réalisiez très prochainement une présentation qui consiste à :

- Concevoir et mettre en prod une DB NoSQL orientée document (MongoDB) ou orientée graphe (Neo4j) sur une plateforme de votre choix ; par ex : Atlas (pour MongoDB) ou AuraDB (pour Neo4j). *N'oublier pas de dénormaliser, d'indexer*, ...

- y injecter un(deux) dataset(s) (complémentaires) : choix du jeu de données (à valider par votre manager/le prof) :

    - kaggle, data.gouv.fr, opendatafrance.net, data-publica.eu, ...

    - Secteur d'activité (de votre employeur d'alternance) : *banque*, *assurance*, *grande distribution*, ...
    - Questions d'actualité via une API et/ou du scrapping pour acquérir des données mises à jour assez régulièrement :

- Spotify-API, twitter-API, Covid, Air France API, SNCF-API, socrata.com crypto-monaie-API, foot-ball-API ... et plus généralement voici un portail des API N'oublier pas de rajouter un champ pour indiquer la date précise

- **CRUD** des données (*langage Python*) : *insertion*, *suppression*, *mise à jour*, *remplacement*,… (manipulation des documents)

- Réaliser un rapport analytique à l'aide de **requêtes d'agrégation et de la viz** pour permettre de répondre aux questions-métier soulevées à partir du jeu de données (questions-métier à valider par le formateur) ; vous pouvez passer par :

  - une WebApp mise en prod sur Heroku par exemple : Flask, Django, Streamlit, ...
  - Tableau / PowerBI

- N'oublier pas d'écrire les scripts d'administration de la DB : *dump*, *restauration*, import de fichier csv/json volumineux, ... , d'améliorer les performances, et de sécuriser la DB.

Livrables attendus (15 pts)

- **Un repo Github** où on trouve :
  - Un fichier README où vous expliquez : **(3 pts)**
    - *Comment lancer les scripts* (de la WebApp, ...)
    - *Mentionnez le lien de votre trello avec la répartition des tâches à travers les membres de votre groupe* (à faire/en cours/fait)
    - …
    - *Lien vers le support de présentation* (par exemple google slides) où on trouve *les principaux résultats* (pour votre manager)
  - - Un notebook Python (non cleané, pour comprendre votre démarche) : **(5 pts)**
    - *Les problèmes rencontrés sur le jeu de données*
    - *Comment vous avez exploré et nettoyé les données*
    - *ETL-Query-Plot qui permettent de répondre aux questions posées*
  - Le code générant la WebApp et permettant de la déployer (par ex sur Heroku) **(7 pts)**

Modalités de présentation du travail (5 pts)

Votre présentation pourra prendre cette forme (à titre indicatif) :

5 min.	Rappel de la problématique, présentation du jeu de données, de l'exploration, du cleaning
15 min	Explication de l'approche pour mettre en place votre solution et démo
10 min	Séance de questions-réponses