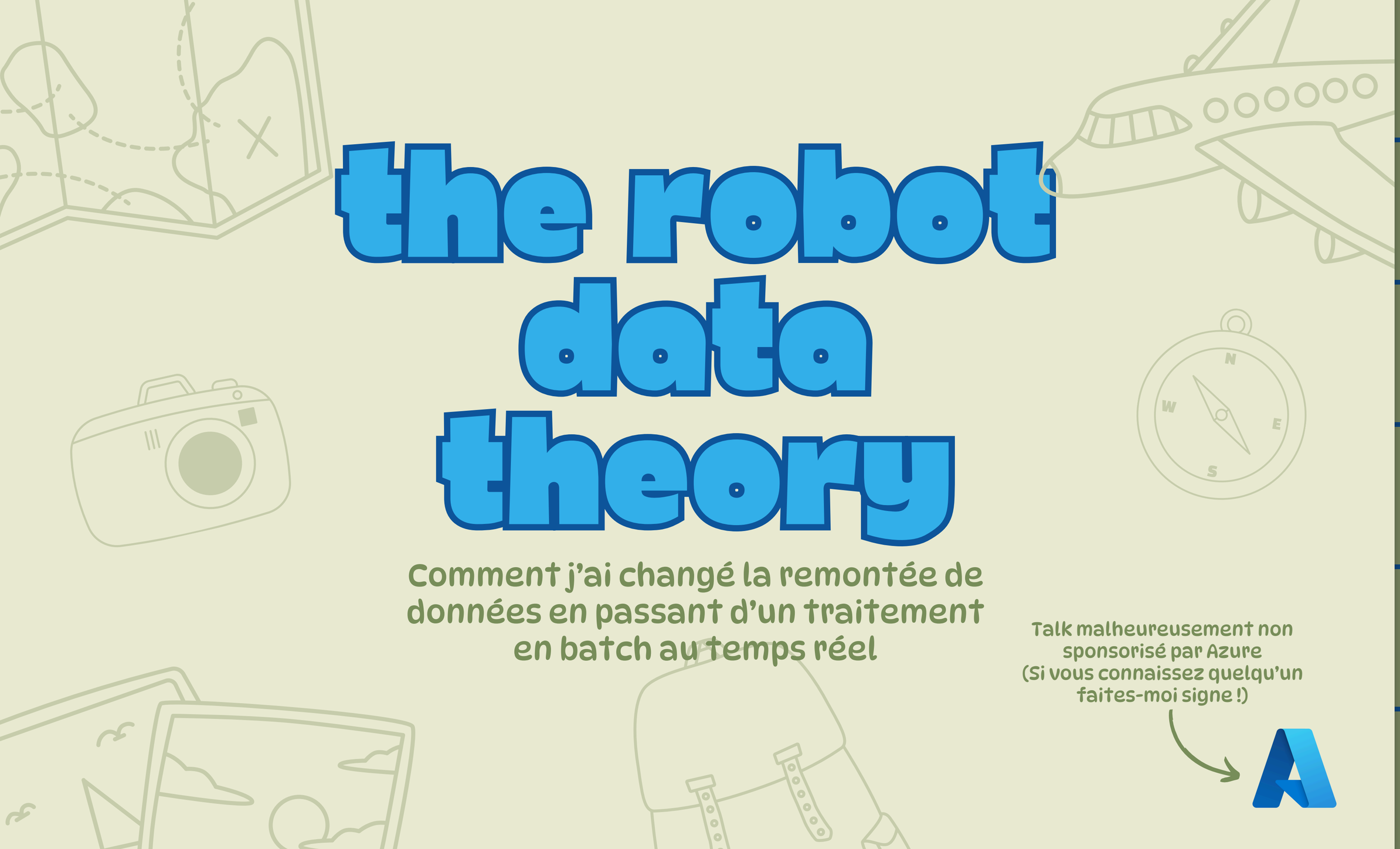
the robot data theory

Comment j'ai changé la remontée de données en passant d'un traitement en batch au temps réel

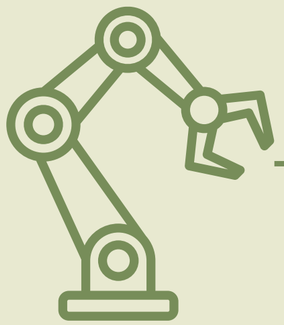
Talk malheureusement non sponsorisé par Azure
(Si vous connaissez quelqu'un faites-moi signe !)



origin story

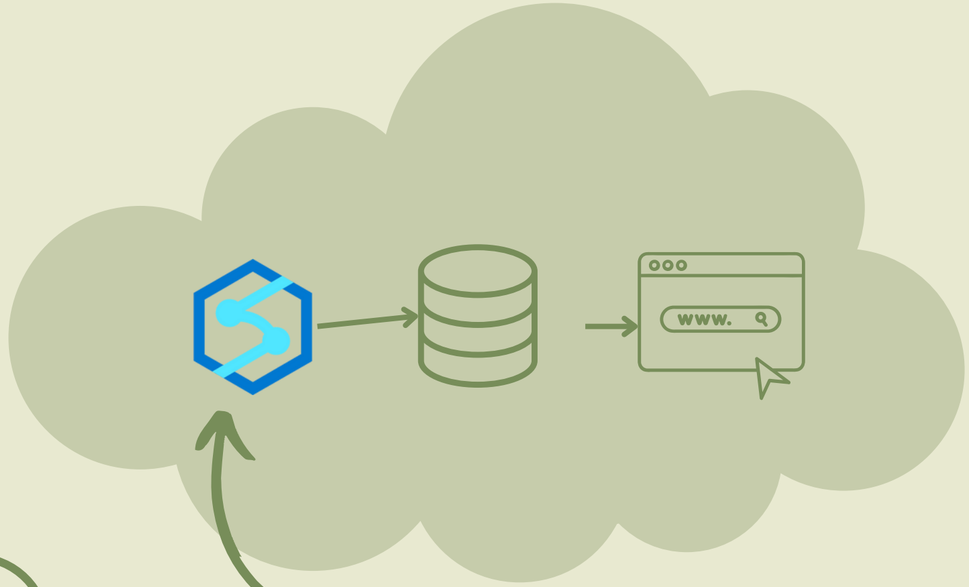
Le début de l'aventure

L'objectif c'est le temps réel, mais bon on en est loin ! Si on arrive à faire moins de 15 minutes ce sera déjà bien !



Nono le robot

Cloud Azure



quelques 4000 lignes de code python

optimisation

Synapse coûte cher : l'outil facture au temps d'exécution du code

Les pipelines de transformation des données sont longues : on recalcule TOUT à chaque fois...

- Analyse du code
- Chargement de moins de tables
- Optimisation de certaines requêtes
- Division du code d'un gros notebook en plusieurs petits
- Parallélisation de certaines tâches

la remise en cause de l'architecture

On utilise peut-être pas le bon outil ?

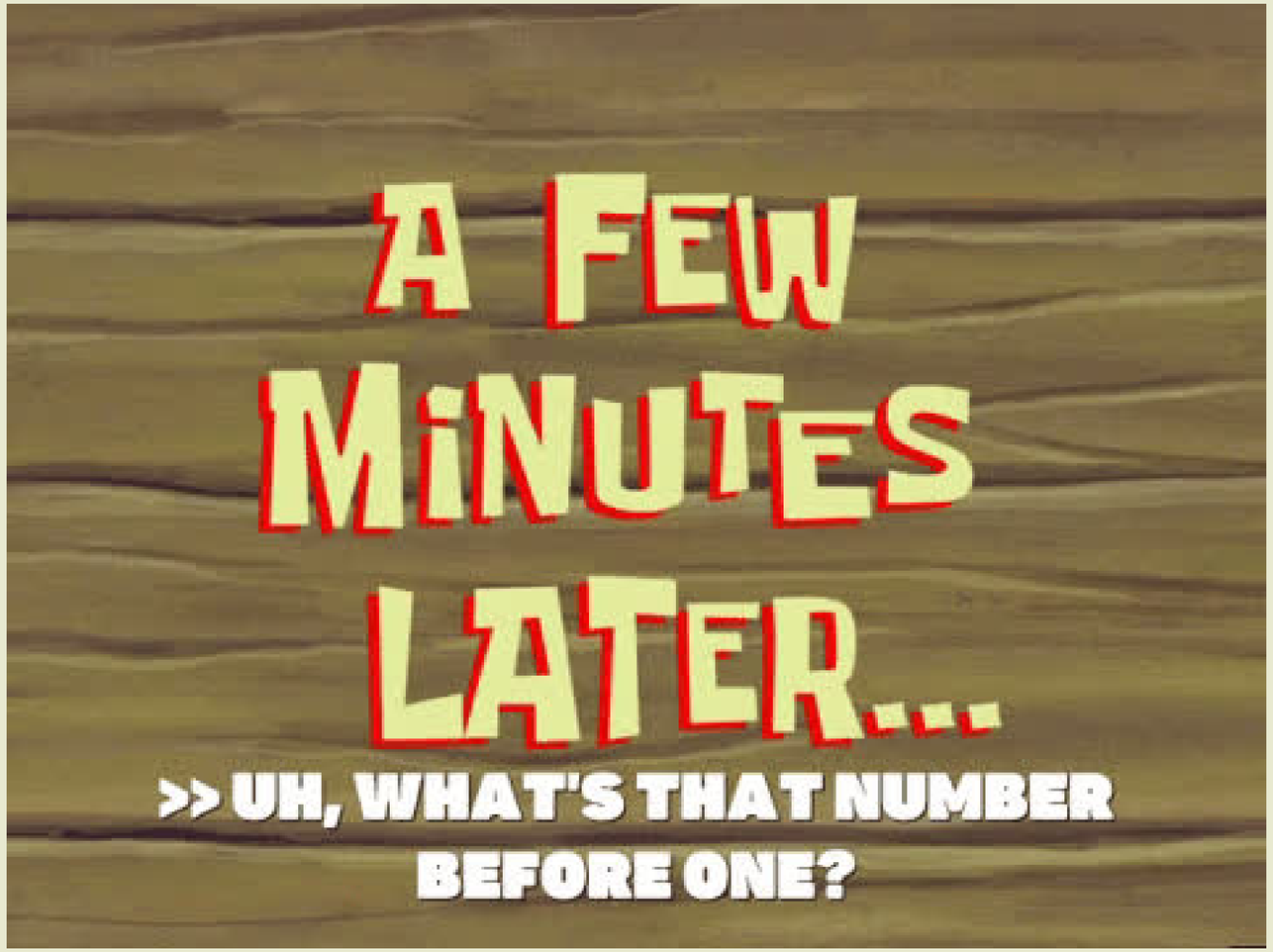
Azure Synapse

- Traitement en **batch**
- Coûte environ 1000€/mois

VS

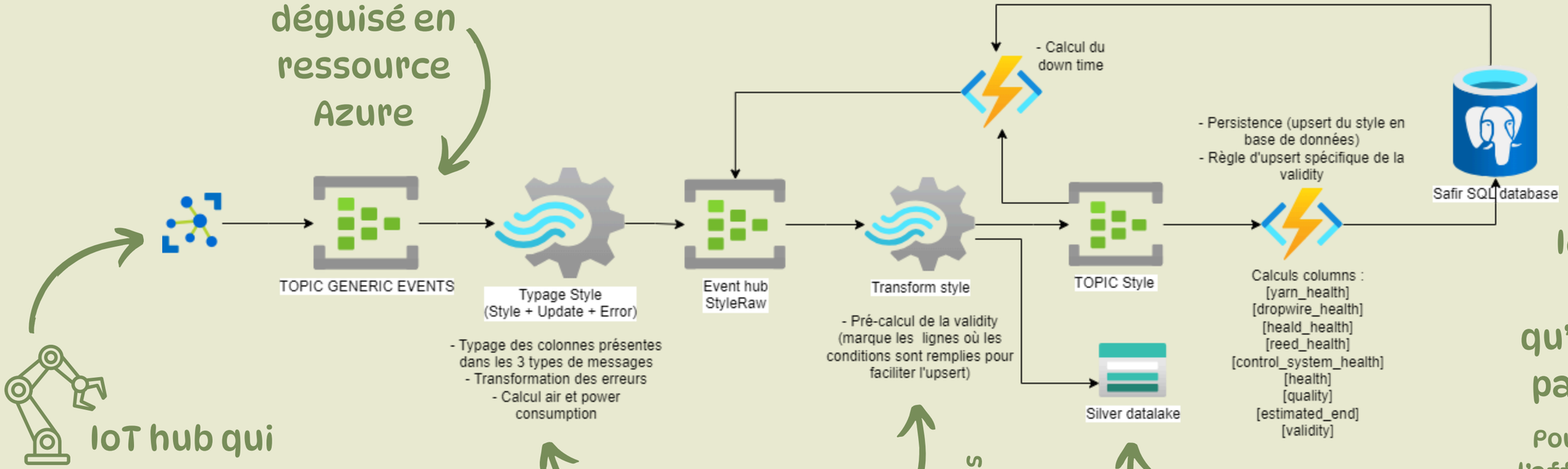
Azure Stream analytics

- Traitement en **stream**
- Coût estimé à 100€/stream/mois


A FEW
MINUTES
LATER...
>> UH, WHAT'S THAT NUMBER
BEFORE ONE?

ce qui a donné

Du Kafka déguisé en ressource Azure



IoT hub qui reçoit les messages de Nono le robot

Typage des colonnes

Ajout de champs calculés

Les transformations de la donnée

Là où on stocke tout

Ici on ne garde qu'une ligne par "style"

Pour optimiser l'affichage sur le dashboard

structurer la donnée

Datalake

Ensemble des données d'un projet (voire de l'entreprise)

La donnée peu être non structurée

La donnée n'est pas (ou peu) transformée

La donnée n'est pas forcément fiable

Base de données

Dédiée à un usage précis

Donnée structurée

La donnée est optimisée pour l'analyse/le reporting

La donnée est historiée

structurer la donnée

Datalake

Data warehouse

Data mart

Base de données

structurer la donnée

Datalake

Data warehouse

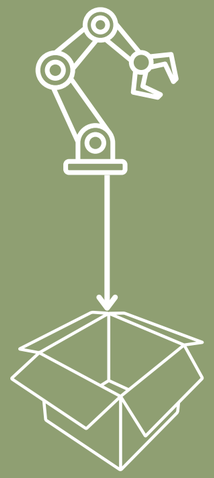
Data mart

La distinction n’est pas clairement faite

Base de données

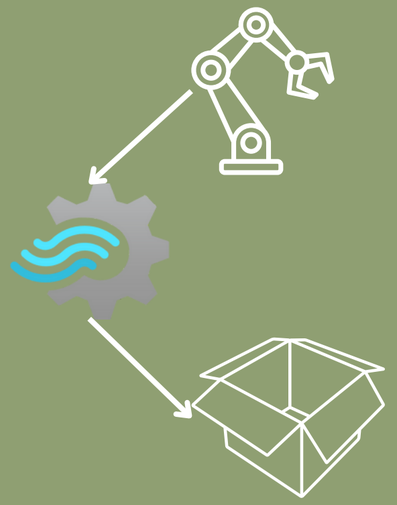
Bronze

Donnée brute



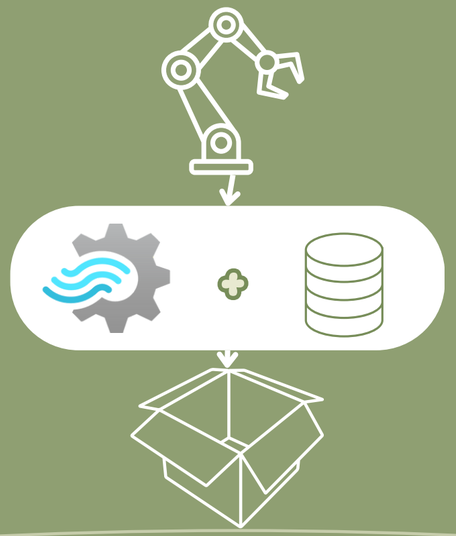
Silver

Donnée validée



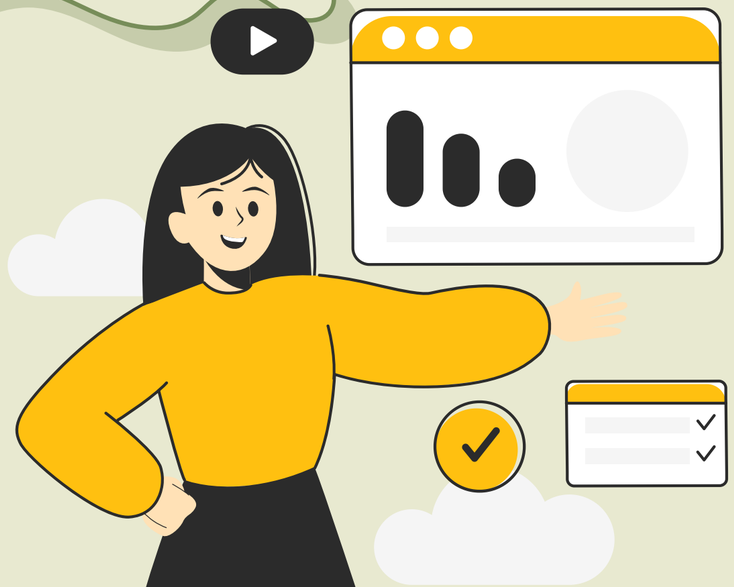
Gold

Donnée enrichie



Bilan :

- Des coûts divisés par 2 (voire 4) pour chaque environnement
- Un temps de traitement qui passe de 45min à moins d'une minute
- Un meilleur accès aux données, que ce soit pour de la manipulation ou du débogage



ce qu'on retient

Analyse du besoin : stream ou batch

Suite à une première analyse erronée (supposer que le client était plus intéressé par l'analyse de données que le monitoring), l'équipe est partie sur le mauvais outil

Bien ranger les données

Base de données, datalake ?
Zone bronze, silver, gold ?
Encore une fois le choix est lié au besoin, et il n'y a pas qu'une seule bonne réponse !