Test

Python et Data Engineering:

1. Data pipeline

Fichiers Repo

2. Traitement ad-hoc

À partir du fichier final on peut trouver le journal avec le plus de citation de drug ce journal, c'est **“The journal of maternal-fetal & neonatal medicine”** avec trois médicaments cités.

3. Pour aller plus loin

Pour être capable de gérer de grosses volumétries de données nous avons besoin de faire évoluer le code, en utilisant des frameworks comme **Spark** pour un traitement distribué sur un cluster, ou on pourra tirer profit le l’écosystème Big data **Hadoop.**

Pour ce faire nous avons besoin d’utiliser **Pyspark,** pour la création de nos Jobs de traitement et une utilisation d’un DAG pour une orchestration optimale.

SQL :

1. Première partie du test

```
Select date, sum(prod_price*prod_qty) as ventes

From TRANSACTIONS

WHERE date BETWEEN '01/01/2019' AND '31/12/2019'

Group by(date)
```

2. **Seconde partie du test**

```
SELECT client_id,sum(prod_price*prod_qty) as ventes

FROM TRANSACTIONS

join PRODUCT_NOMENCLATURE

On TRANSACTIONS.prod_id= PRODUCT_NOMENCLATURE.product_id

WHERE date BETWEEN '01/01/2019' AND '31/12/2019'

group by client_id,product_type
```