Approches sémantiques pour l'exploration d'un lac de donnée

Projet M2 DataScale
Encadré par : Zoubida Kedad

Novembre 2025

1

Contexte : les lacs de données

Data Warehouse **Vs.** **Data Lake**

Nombre limité de sources de données.
Stockage après traitement
Moins flexible
Moins évolutif (coûteux)

- *Grands volumes de données hétérogènes, brutes*
- *Stockage direct*
- *Plus flexible*
- *Très évolutif*

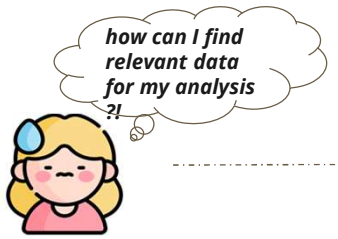
MAIS :

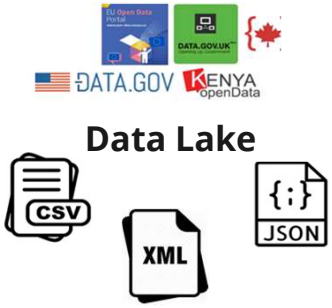
- *Absence de schéma.*
- *Métadonnées peu fiables, incohérentes et incomplètes.*

2

Exploration d'un lac de données

Problème : retrouver les données pertinentes pour une tâche d'analyse spécifique.





- ✖ Collection de sources de données brutes.
- ✖ Absence de schéma.
- ✖ Métadonnées peu fiables, incohérentes et incomplètes.

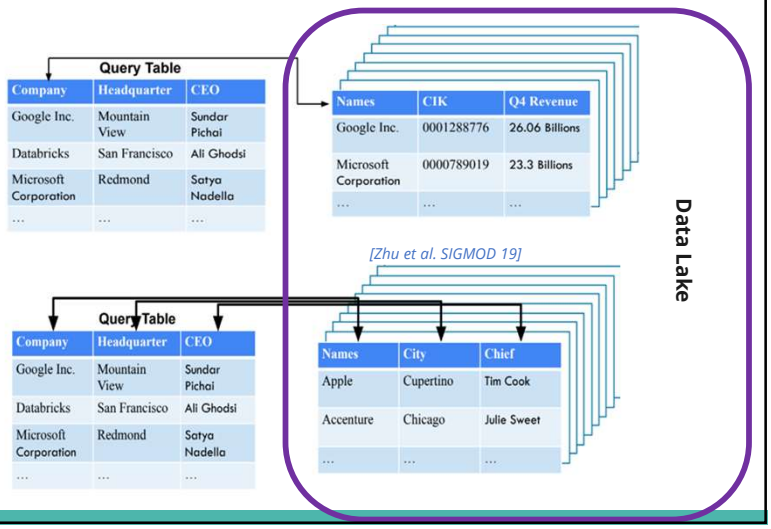
Comment faciliter l'exploration et la recherche des données ?

Recherche de tables pertinentes

Problème : étant données une table requête T soumise par l'utilisateur, comment retrouver les tables du lac de données qui peuvent être combinées par jointure ou par union avec T ?

Company	Headquarter	CEO
Google Inc.	Mountain View	Sundar Pichai
Databricks	San Francisco	Ali Ghodsi
Microsoft Corporation	Redmond	Satya Nadella
…	…	…

Query Table



Recherche de tables pertinentes

Principe :

- Générer des annotations pour chaque table du lac de données.
- Lorsque l'utilisateur soumet une table requête, cette dernière est annotée également.
- La recherche des tables est faite en comparant les annotation de la table requête et les annotations des tables du lac de données.

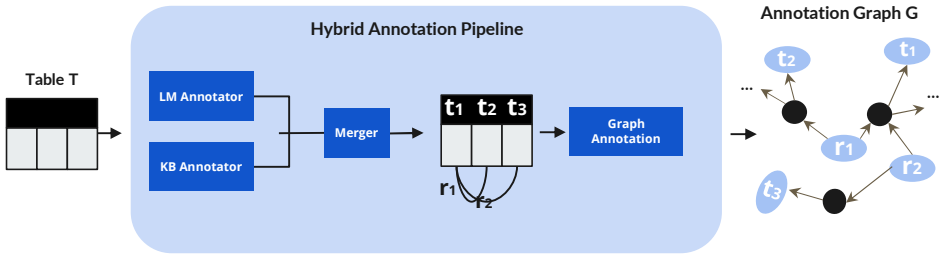
Le projet se focalise sur deux tâches (décrites dans les 2 slides suivants) :

1. Annotation de tables : concevoir et implémenter une approche qui génère un graphe d'annotation pour une table, en utilisant des modèles de langages et des bases de connaissances en ligne.
2. Recherche de tables : définir et implémenter un processus de recherche qui retrouve des tables pertinentes à partir d'une table requête en comparant leurs graphes d'annotation.

5

Tâche 1 : Annotation des tables d'un lac de données.

Proposer une approche d'annotation hybride qui combine l'utilisation de modèles de langages et de bases de connaissances.

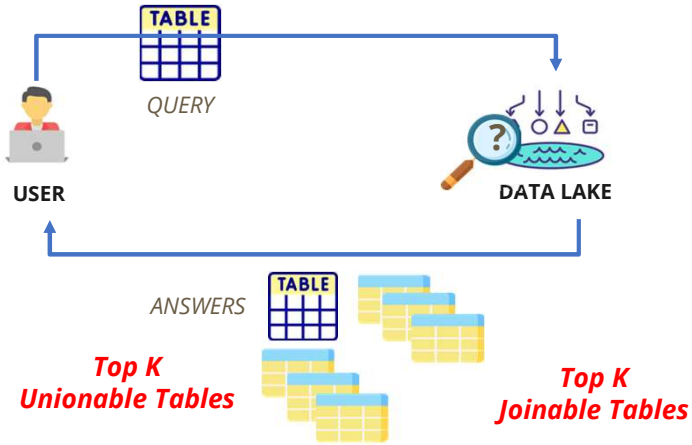


- Fine-tuner des modèles de langages sur la tâche de découverte de types et de relations :
  - en testant different LMs
  - avec différent façons de serialiser les tables

6

Tâche 2 : Recherche de tables dans un lac de données

- Rechercher des tables qui peuvent compléter une table requête par union (ajout de tuples) ou par jointure (ajout de colonnes), en définissant un processus de matching de graphes d'annotation.



7

Travail à réaliser

- Etude de 2 ou 3 approches d'annotation et de recherche dans des lacs de données.
- Définir et implémenter des approches d'annotation et de recherche.
- Evaluation des approches en utilisant un benchmark existant.

- Référence bibliographique Doduo : https://arxiv.org/pdf/2104.01785

8