Sujet M1 - UCBL - Ouverture à la recherche (pour un groupe de 3 étudiants)

Conversion and exploration de Données PFAS du CNRS

Prof. Angela Bonifati (Liris, UCBL, IUF - France)
https://perso.liris.cnrs.fr/angela.bonifati/

Les jeux de données PFAS data hub (https://pdh.cnrs.fr/en/) contiennent les informations sur les PFAS (polluants éternels). Cette collecte de données est issue d'un projet de recherche (The Forever Pollution project) et contient 104 jeux de données en format CSV et parquet.

Ces jeux de données ont une provenance différentes et variées selon la carte: https://pdh.cnrs.fr/en/map/

Leur intégration et exploration dans le format actuel est très ardue à cause du fait que chaque dataset possède ses propriétés (attributs dans le CSV).

Le travail de ce projet M1 par 2 voir 3 étudiants consiste à:

- Envisager une transformation de jeux de données dans un graphe de propriété, un graphe avec plusieurs labels sur les noeuds/ arrêtés et une liste de clé/valeur sur ces derniers;

- L'éxtraction de schémas de différents datasets avec des outils de property graph schema discovery, comme DiscoPG (code accès libre sur github) https://www.vldb.org/pvldb/vol15/p3654-bonifati.pdf

- L'intégration et appariement de ces schémas et leur conversion en PG-Schema (un parseur de PG-Schema sera mis à disposition - code en accès libre sur zenodo https://zenodo.org/records/7362078
-

- Exploration de schémas et de jeux de données pour comprendre certains pattern récurrents, corrélations et liens de causalité

Réferences

[DiscoPG github] https://github.com/PI-Clustering/code

[ABD23] Renzo Angles, Angela Bonifati, Stefania Dumbrava, George Fletcher, Alastair Green, Jan Hidders, Bei Li, Leonid Libkin, Victor Marsault, Wim Martens, Filip Murlak, Stefan Plantikow, Ognjen Savkovic, Michael Schmidt, Juan Sequeda, Slawek Staworko, Dominik Tomaszuk, Hannes Voigt, Domagoj Vrgoc, Mingxi Wu, Dusan Zivkovic: PG-Schema: Schemas for Property Graphs. Proc. ACM Manag. Data 1(2): 198:1-198:25 (2023)

[BDM22] Angela Bonifati, Stefania-Gabriela Dumbrava, Emile Martinez, Fatemeh Ghasemi, Malo Jaffré, Pacome Luton, Thomas Pickles:
DiscoPG: Property Graph Schema Discovery and Exploration. Proc. VLDB Endow. 15(12): 3654-3657 (2022)