ÉLÉMENT DE PORTFOLIO 01

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément :

Discovering and merging related analytic datasets

URL de l'élément : `https://hal.sorbonne-universite.fr/hal-02459098v1`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Cet article est le résultat d'une première collaboration importante avec la société SAP France dans le cadre de la thèse CIFRE de Rutian Liu [3] et une publication dans le journal Information Systems [4]. Les résultats pratiques de ce travail ont été développés avec le logiciel SAP HANA pour ensuite être intégré dans le produit SAP Data Intelligence. Cette collaboration a permis de continuer une collaboration sur la qualité de requêtes analytiques [5] et sur la parallélisation de data pipelines dans Data Intelligence (un stage M2 en 2022 et deux stages M2 en cours en 2023).

3 PRÉSENTATION DE CET ÉLÉMENT

La croissance importante des sources de données générées rend leur utilisation de plus en plus complexe. Par exemple, la société SAP fournit à ces clients des milliers de jeux de données analytiques prédéfinis et personnalisables pour divers domaines d'applications industrielles et commerciales (gestion de la chaîne logistique, gestion de la relation client, planification de ressources d'entreprises). Ces jeux de données sont définis comme des vues analytiques (multidimensionnelles) sur les données transactionnelles stockées et gérées par la suite SAP S4/HANA et contiennent des informations avec des mesures sophistiquées prêtes à être utilisées par les utilisateurs.

Des outils de préparation des données permettent aux utilisateurs de créer leurs propres ensembles de données analytiques de haute qualité à partir des données transactionnelles et des données analytiques existants, et de créer des visualisations de données personnalisées interactives et perspicaces. Un besoin typique des utilisateurs de ces outils est d'augmenter le schéma d'une table de données analytiques existant avec de nouveaux attributs, provenant d'un ou plusieurs jeux de données sémantiquement liés, qui peuvent représenter des détails supplémentaires sur les dimensions ou de nouvelles mesures. Il s'agit d'un besoin essentiel dans de nombreux scénarios tels que la création de tableaux de bord ou la préparation de données pour apprendre des modèles prédictifs. Malgré l'importance de cette tâche, les outils actuels offrent peu d'aide aux utilisateurs pour créer de nouvelles tables analytiques. Ce manque de soutien conduit les utilisateurs à redéfinir à plusieurs reprises des ensembles de données analytiques similaires d'une manière parfois incohérente ou à faire appel à leur service informatique pour créer leurs ensembles de données personnalisés.

Cet article étend les solutions existantes sur l'augmentation de tables relationnelles [1, 2, 6] au cas de tables de données analytiques. Le défi principal dans l'augmentation du schéma d'une table analytique est de respecter des contraintes imposées par les dimensions et les mesures présentes dans les tables utilisées et les propriétés des fonctions d'agrégation. Ainsi, une requête de jointure simple peut introduire des incohérences dans la table augmentée en produisant plusieurs lignes par clé (row multiplication) et des mesures agrégées incorrectes ou ambigües (incorrect / ambiguous reduction). L'article présente un modèle d'augmentation de données analytiques avec des critères de qualité formels pour les tables augmentées par des jointures avec d'autres tables analytiques. Ces critères sont e utilisés (1) pour identifier automatiquement des opérations de réparation, (b) pour notifier aux utilisateurs la génération d'attributs ambigus, (c) pour déduire des fonctions d'agrégation applicables sur les nouveaux attributs, et (d) pour compléter les résultats de fusion obtenus par une augmentation de schéma incomplète. La solution proposée inclut une analyse théorique minutieuse et une implémentation dans le système SAP HANA.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy, Hongrae Lee, Fei Wu, Reynold Xin, and Cong Yu. Finding Related Tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 817–828, New York, NY, USA, 2012. ACM.

[2] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, and Nan Tang. The Data Civilizer System. In *CIDR*, 2017.

[3] Rutian Liu. *Semantic Services for Assisting Users to Augment Data in the Context of Analytic Data Sources*. Theses, Sorbonne Université, June 2020.

[4] Rutian Liu, Eric Simon, Bernd Amann, and Stéphane Gançarski. Discovering and merging related analytic datasets. *Information Systems*, 91 :101495, July 2020.

[5] Eric Simon, Bernd Amann, Rutian Liu, and Stéphane Gançarski. Controlling the Correctness of Aggregation Operations During Sessions of Interactive Analytic Queries. *ACM Journal of Data and Information Quality*, January 2023.

[6] Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. InfoGather : Entity augmentation and attribute discovery by holistic matching with web tables. In *Proceedings of the 2012 International Conference on Management of Data - SIGMOD '12*, pages 97–108. ACM Press, 2012.