

La modélisation ensembliste au banc d'essais

jerome.darmont@univ-lyon2.fr

Contexte

La modélisation ensembliste (*ensemble modeling*) est une approche dont l'objectif est de rendre plus facile l'évolution des schémas relationnels des entrepôts de données (Rönnbäck & Hultgren, 2013). Il existe deux approches de modélisation ensembliste : les *data vaults* (Hultgren, 2012 ; Linstedt, 2015 ; Eshetu, 2014) et l'*anchor modeling* (Regardt et al., 2009 ; Rönnbäck et al., 2010). Elles sont en fait très proches (Rönnbäck & Hultgren, 2013), avec une évolutivité un peu plus facile et une évolution de schéma non destructive pour l'*anchor modeling*, mais un plus grand nombre d'objets à gérer en raison d'une modélisation en sixième forme normale (6NF), ainsi que des procédures de maintenance des attributs temporels (*timestamps*) non automatisées. Les *data vaults* sont plus proches de la modélisation multidimensionnelle traditionnelle (ils adoptent la troisième forme normale – 3NF) et sont supportés par un plus grand nombre d'outils. Bien que décrits uniquement au niveau logique (relationnel), les concepts de la modélisation ensembliste peuvent facilement être transposés aux niveaux conceptuel (Jovanovic & Bojicic, 2012) et physique (Krnetta et al., 2014).

Objectif

La modélisation ensembliste induit des classes (au niveau conceptuel) et des tables (niveaux logique et physique) plus nombreuses que dans une modélisation en étoile classique. Cela implique également des jointures plus nombreuses lors des interrogations de l'entrepôt de données. Toutefois, les auteurs de l'*anchor modeling* pensent que la modélisation en 6NF permet de bons temps de réponse, grâce à la technique d'élimination de jointures employée dans les optimiseurs de requêtes modernes (Paulley, 2000).

Afin de nous en assurer, l'objectif de ce TER est de comparer une modélisation multidimensionnelle classique d'entrepôt de données à sa modélisation en *data vault* et par *anchor modeling*. Pour faire cette comparaison expérimentale, nous allons utiliser un banc d'essais constitué d'un entrepôt de données test, d'une charge de requêtes appliquée à l'entrepôt, d'un protocole d'exécution et de mesures de performances (Darmont, 2017). Dans le cadre de ce travail, il s'agit de mettre en œuvre le banc d'essais Star Schema Benchmark (SSB ; O'Neil et al., 2009).

Tâches à effectuer

1. Lire les documents concernant les *data vaults*, l'*anchor modeling* et SSB cités en références bibliographiques. En trouver d'autres si possible.
2. Traduire le modèle logique relationnel de l'entrepôt de données de SSB en *data vault* et en *anchor model*.
3. Traduire les trois modèles logiques (SSB en étoile, en *data vault* et en *anchor model*) en un modèle physique PostgreSQL¹.
4. Adapter la charge de requêtes SQL de SSB au *data vault* et à l'*anchor model*.
5. Exécuter le banc d'essais avec les trois modèles d'entrepôt, les requêtes SSB adaptées aux trois modèles et différents facteurs d'échelle (*scale factor* ou SF dans SSB). Mesurer le temps d'exécution de la charge de requêtes pour chaque expérience.
6. Écrire un article d'une dizaine de page maximum, de préférence à l'aide de LaTeX², en français ou en anglais, qui présente votre travail. Plan de l'article :
 - a. Introduction (contexte, problématique, contribution, annonce du plan de l'article)
 - b. État de l'art : modélisation ensembliste et bancs d'essais décisionnels (OLAP Council, 1998 ; Darmont et al., 2005 ; TPC 2017, 2018)
 - c. Adaptation de SSB en *data vault* et en *anchor model* (modèle de données et charge de requêtes)

¹ <https://www.postgresql.org/>

² <https://www.latex-project.org>

- d. Comparaison expérimentale (conditions expérimentales, présentation des expériences, résultats/discussion)
- e. Conclusion

Références bibliographiques

- Darmont J., Bentayeb F., Boussaïd O. (2005). DWEB: A Data Warehouse Engineering Benchmark. 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2005), Copenhagen, Denmark. LNCS, 3589, 85-94. <https://hal.archives-ouvertes.fr/hal-00145452/document>
- Darmont J. (2017). Data-Centric Benchmarking. Encyclopedia of Information Science and Technology, 4th Edition. IGI Global. 1772-1782. <https://hal.archives-ouvertes.fr/hal-01547328/document>
- Eshetu F.A. (2014). Data Vault Modelling: An Introductory Guide. BSc thesis, Helsinki Metropolia University of Applied Sciences. <https://www.theseus.fi/handle/10024/74895>
- Hultgren H. (2012). Data vault modelling guide – Introductory guide to data vault modelling. Genesee Academy. <https://hanshultgren.files.wordpress.com/2012/09/data-vault-modeling-guide.pdf>
- Jovanovic V., Bojicic I. (2012). Conceptual Data Vault Model. Southern Association for Information Systems Conference, Atlanta, GA, USA: 131-136. <https://works.bepress.com/vladan-jovanovic/9/>
- Krneta D., Jovanovic V., Marjanovic Z. (2014). A Direct Approach to Physical Data Vault Design. Computer Science and Information Systems, 11(2): 569-599. <http://www.comsis.org/pdf.php?id=472-1305>
- Linstedt D. (2015). Data Vault Basics. <https://danlinstedt.com/solutions-2/data-vault-basics/>
- OLAP Council. (1998). APB-1 OLAP Benchmark Release II. <http://www.olapcouncil.org/research/bmarkly.htm>
- O’Neil P., O’Neil E., Chen X., Revilak, S. (2009). The Star Schema Benchmark and Augmented Fact Table Indexing. 1st Technology Conference on Performance Evaluation and Benchmarking (TPCTC 2009), Lyon, France. LNCS, 5895, 237-246. <https://www.cs.umb.edu/~poneil/StarSchemaB.PDF>³
- Paulley, G.N. (2000). Exploiting Functional Dependence in Query Optimization. PhD thesis, University of Waterloo, Canada. <https://uwspace.uwaterloo.ca/bitstream/handle/10012/511/NQ51220.pdf?sequence=1>
- Regardt O., Rönnbäck L., Bergholtz M., Johannesson P., Wohed P. (2009). Anchor Modeling. 28th International Conference on Conceptual Modeling (ER 2009), Gramado, Brazil. LNCS, 5829: 234-250. https://www.researchgate.net/publication/221268907_Anchor_Modeling
- Rönnbäck L., Regardt O., Bergholtz M., Johannesson P., Wohed P. (2010). Anchor Modeling – Agile information modeling in evolving data environments. Data and Knowledge Engineering, 69(12): 1229-1253. <http://www.anchor modeling.com/wp-content/uploads/2011/05/Anchor-Modeling.pdf>
- Rönnbäck L., Hultgren, H. (2013). Comparing Anchor Modeling with Data Vault Modeling. https://hanshultgren.files.wordpress.com/2013/06/modeling_compare_05_larshans.pdf
- TPC. (2017). TPC Benchmark H Standard Specification Version 2.17.3. Transaction Processing Performance Council. http://www.tpc.org/tpc_documents_current_versions/pdf/tpc-h_v2.17.3.pdf
- TPC. (2018). TPC Benchmark DS Standard Specification Version 2.10.0. Transaction Processing Performance Council. http://www.tpc.org/tpc_documents_current_versions/pdf/tpc-ds_v2.10.0.pdf

³ Voir aussi le générateur de données de SSB, dbgen : <http://www.cs.umb.edu/~poneil/dbgen.zip>