Fonctions d'oubli dans les entrepôts de données

Aliou Boly*, **, Georges Hébrail*, Marie-Luce Picard**

*Ecole Nationale Supérieure des Télécommunications de Paris
46, Rue Barrault 75634 Paris cedex 13 France
boly@enst.fr, hebrail@enst.fr
**Electricité de France Recherche et Développement
1, Av. du Général de Gaulle 92141 Clamart Cedex
marie-luce.picard@edf.fr

Résumé. Les entrepôts de données stockent des quantités de données de plus en plus massives, en particulier du fait de la constitution d'historiques. Nous proposons ici une solution pour éviter la saturation des entrepôts de données. Nous définissons un langage de spécifications de fonctions d'oubli des données les plus anciennes, permettant de déterminer ce qui doit être présent dans l'entrepôt de données à chaque instant. Ces spécifications de fonctions d'oubli se traduisent par des opérations de résumé par agrégation, et par des opérations de suppression des données anciennes réalisées de façon mécanique à chaque pas de mise à jour. La communication présente tout d'abord une description syntaxique du langage de spécifications des fonctions d'oubli. Les contraintes à vérifier pour assurer la cohérence du langage sont ensuite décrites. Enfin, nous proposons des structures de données adaptées au stockage des données nécessaires à la gestion des fonctions d'oubli.

1 Introduction

L'objectif de cette communication est d'apporter une réponse au problème de saturation des entrepôts de données, en définissant un langage de spécifications de fonctions d'oubli des données. Ces spécifications conduisent à supprimer de façon mécanique les données à 'oublier', tout en conservant un résumé de celles-ci par agrégation. Pour définir ces stratégies d'oubli des données, le langage considère comme critère d'oubli une dimension temporelle : l'ancienneté de la donnée. L'ancienneté d'une donnée, définie au niveau du n-uplet d'une table, peut être soit la date de dernière mise à jour du n-uplet (fournie par le SGBD et appelée timestamp), soit de façon explicite par une colonne de type Date. Les spécifications d'oubli sont définies sur chaque table, prises indépendamment les unes des autres dans ce travail.

Ce travail est à relier aux travaux sur les bases de données temporelles [Dumas *et al.,* 2001] où l'on cherche à dater les informations et à gérer leur historique, et aux travaux sur la gestion des versions dans les bases de données [Cellary *et al,.* 2001], où l'on cherche à conserver et manipuler les évolutions de données. Il peut également être relié aux travaux de [Chaudhuri *et al.,* 1998] et [Chaudhuri *et al.,* 2001], qui cherchent à estimer les résultats de requêtes sur une table volumineuse en ne considérant qu'un échantillon du contenu de la table. Cependant, notre problématique est différente car dans ces travaux, l'ensemble des données reste toujours en ligne, alors que nous proposons ici des stratégies pour archiver définitivement les données détaillées anciennes, en n'en conservant qu'une version agrégée.