REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA
FACULTE DES MATHEMATIQUES ET DE L'INFORMATIQUE

DEPARTEMENT D'INFORMATIQUE

MEMOIRE de fin d'étude

Présenté pour l'obtention du diplôme de MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Systèmes d'Informations Avancés

Par : MOUSSAOUI Rime

SUJET

Métaheuristiques appliquées au problème de fragmentation d'entrepôts de données : Etude comparative

Soutenu publiquement le : 01 /06 /2016 devant le jury composé de :

Dr. LAMICHE Chaabane	**Université de M'sila**	**Rapporteur**
N. Mouhoub	**Université de M'sila**	**Président**
A. Khettaf	**Université de M'sila**	**Examinateur**
.....	**Université de M'sila**	**Examinateur**

Promotion : 2015 /20 16

Tables des Matières

Chapitre 02 : Optimisation dans les entrepôts de données

Chapitre 03 : Résolution du problème de sélection de schéma de fragmentation optimale

Chapitre 04 : Réalisation et expérimentations

Introduction générale

Les entrepôts de données (data warehouse en anglais) c'est un outil décisionnel qui permet l'exploitation des données d'une organisation dans le but de faciliter la prise de décision. Pour cela les dirigeants des entreprises intégrant des données dites de production dans une base de données centralisée (entrepôts) ou elles sont agrégée, historisées et structurée de manière à en permettre le regroupement, nettoyage et intégration des données et aussi établissement des requêtes, rapports et des analyses et offre la possibilité d'extraire des connaissances (fouille de données).

Les entrepôts de données sont souvent modélisés par un schéma en étoile. Cette dernierre est caractérisé par une table de fait de très grand taille et un ensemble de tables de dimensions de plus petite taille.

Le table de fait contient qu'un ensemble de mesures collectées durant l'activité de l'organisation. Les tables de dimension contiennent des données qualitatives qui représentent des axes sur lesquels les mesures ont été collectées. Les requêtes de type OLAP définies sur un schéma en étoile (connues par requêtes de jointure en étoile) sont caractérisées par des opérations de sélection sur les tables de dimension, suivies de jointures avec la table des faits. Aucune jointure n'existe entre les tables de dimension. Toute jointure doit passer par la table des faits, ce qui rend le coût d'exécution de ces requêtes très important Sans technique d'optimisation, leur exécution peut prendre des heures, voire des jours. [1]

Dans le domaine optimisation dans entrepôts de données il existe beaucoup des techniques. Les techniques redondante(les vues matérialisé, les indexes, fragmentation verticale) et techniques non redondante (fragmentation horizontale). Fragmentation horizontale consiste à partitionner la table de fait en un ensemble des fragments en utilisant des fragments des tables de dimensions.

L'objectif de notre étude est l'adaptation du metaheuristique à population (un algorithme génétique) et méthode descente pour obtention un schéma optimal des tables de dimensions permettant de nous donner un schéma de fragmentation optimal d'entrepôts.

Nous avons organisé notre mémoire de la façon suivant :

Dans le chapitre 1, nous avons présenté les éléments fondamentaux d'entrepôts de données. Le chapitre 2 est réservé à la fragmentation des entrepôts de données. La méthodologie proposée

pour la résolution du problème de sélection de schéma de fragmentation optimale est largement discutée dans le chapitre 3. Enfin la réalisation de notre application et nos et expérimentations. sont présentées dans le chapitre 4. Le mémoire se termine par une conclusion où nous avons adressé quelques perspectives.

Conclusion générale

Dans le cadre de ce travail de master, nous avons traité le problème de sélection d'un schéma de fragmentation optimale. Au cœur de ce mémoire nous avons présenté les différentes étapes de la modélisation de la solution, pour mieux cerner la problématique posée, nous avons commencé par la présentation générale des notions concerne les entrepôts de données et une citation des techniques d'optimisation des requêtes. En deuxième temps, nous avons passé à l'explication en détail de la technique d'optimisation basée sur la fragmentation horizontale. Pour guider notre choix de résolution, nous avons présenté d'une manière générale les algorithmes génétiques et ces opérateurs génétiques tels que codage, mutation et croisement et la méthode descente puis, nous avons adopté ces méthodes pour résoudre le problème de sélection d'un schéma de fragmentation optimale.

Nous avons choisi le langage c# pour écrire et développé notre application, on a utilisé oracle 10g pour la création de notre Schéma en étoile et on a utilisé les données fournis par banc essai benchmark.

Comme perspective à ce travail, l'utilisation d'autres métaheuristiques hybrides est désirée afin de trouver une solution approchée au problème de sélection d'un schéma de fragmentation optimale. On peut aussi augmenter la charge de requêtes utilisée pour montrer la performance de notre application. Enfin, et pour évaluer la technique de fragmentation développée, il est important d'utiliser le schéma générer par l'application pour fragmenter l'entrepôt de données ce qui réduit le temps d'exécution des requêtes.

En effet, ce travail étant une œuvre humaine n'est pas une solution unique et parfait, c'est pourquoi nous restons ouverts à toutes les critiques et nous somme prêts à recevoir toutes suggestions.

Bibliographie

[1] E. Ziyati, « optimisation de requêtes olap en entrepôts de données approche basée sur la fragmentation génétique », thèse de doctorat, université mohammed v –agdal rabat – maroc, mai 2010.

[2] http://www.upmf-grenoble.fr/ , consulté : le 05/04/20016

[3] D. Garar, « compression dans les entrepôts de données pour l'amélioration des performances », thèse de doctorat, université du québec à montréal, janvier 2013.

[4] B. Espinasse, « introduction aux entrepôts de données », université de marseille, septembre 2013.

[5] K. Boukhalfa, « de la conception physique aux outils d'administration et de tuning des entrepôts de données », thèse de doctorat, école doctorale : sciences pour l'ingénieur et aéronautique, juillet 2009.

[6] N. Zanoun, « la construction en ligne des tables individus variables par apprentissage automatique numérique (pmc) », mémoire de magister, université d'oran, 2010.

[7] A. Daâs ,« optimisation des requêtes dans le data warehouse », mémoire de magister, université ferhat abbas – sétif, 2012.

[8] E. F. codd , providing olap(on-line analytical processing)to user analystes : an it mandate, technical report,1993.

[9] O. teste, modélisation et manipulation d'entrepôts de données complexes et historisées, université paul sabatier - toulouse iii, 2000.

[10] L. Bellatreche, « la conception physique des data warehouses », ladjel bellatreche, lisi/ensma ,2006.

[11] H. jagadish, l. v. s. lakshmanan, and d. srivastava. «what can hierarchies do for your data warehouses», proceedings of the international conference on very large databases, pages 530–541, september 1999.

[12] R. Bouchakri « une approche dirigée par la classification des attributs pour

fragmenter et indexer des entrepôts de données », mémoire de magister, école nationale supérieure d'informatique (esi), 2009.

[13] The olap report - http://www.olapreport.com/fasmi.htm, 2004.

[14] M. trinidad et S. encinas, « entrepôts de données pour l'aide à la décision médicale : conception et expérimentation », maría trinidad serna encinas, université joseph fourier, le 27 juin 2005.

[16] P. vassiliadis and Timos k. sellis, « a survey of logical models for olap databases», sigmod record, 28(4): pp 64–69, 1999.

[17] H. gupta and I. s. mumick, « selection of views to materialize in a data warehouse », ieee trans. on knowledge and data eng, 17(1): pages 24–43, january 2005.

[18] P. valduriez. join indices, «acm transactions on database systems», 12(2) :218–246, june 1987.

[19] Red breck systems , «star schema processing for complex queries». white paper. Juillet 1997.

[20] S. chaudhuri et v. narasayya, «index merging», proceedings of the international conference on data engineering (icde) », pages 296-303, march 1999.

[21] H. mahboubi, « optimisation de la performance des entrepôts de données xml par fragmentation et répartition », thèse de doctorat, université lumière lyon 2, 08/12/2008.

[22] M. barr, « approche dirigée par les fourmis pour la fragmentation horizontale des entrepôts de données relationnels », thèse de magister, e.s.i, 2008.

[23] L. bellatreche et K. boukhalfa, « sélection de schéma de fragmentation horizontale dans les entrepôts de données formalisation et algorithmes », rapport de recherche, lisi/ensma poitiers, université de laghouat, algérie 2006.

[24] M. t. özsu et P. valduriez. « principles of distributed database systems» ,second edition,prentice hall, 1999.

[25] L. bellatreche, K. boukhalfa, et H. i. abdalla. saga « a combination of genetic and simulated annealing algorithms for physical data warehouse design»,in 23rd british national conference on databases, (212-219), july 2006.

[26] R. bouchakri, « conception physique statique et dynamique des entrepôts de données», thèse de doctorat, école nationale supérieure d'informatique (algérie) et école nationale supérieure de mécanique et d'aérotechnique (france),17 septembre 2015.

[27] T. Vallé et M. Yıldızo˘glu, « Présentation des algorithmes génétiques et de leurs applications en Economie», Université Montesquieu Bordeaux IV ,7 septembre 2001.

[28] S. Krour, « optimisation des paramètres d'une cellule photovoltaïque par les algorithmes génétiques », mémoire de magister, université de Farhat abbas 1,21/12/2014.

[29] N. Talbi, « Conception des Systèmes d'Inférence Floue par des Approches Hybrides : Application pour la Commande et la Modélisation des Systèmes Nonlinéaires», thèse de doctorat, Université de Constantine 1, le 25 /02 / 2014.

[30] A. Nafi, « la programmation pluriannuelle du renouvellement des réseaux d'eau potable », thèse de doctorat, université Louis Pasteur, Strasbourg I, 04/12/2006.

[31] F. Souam ait elhadj, « Approche de détection de communautés chevauchantes dans réseaux bipartis », thèse de doctorat, université mouloud Mammeri Tizi-Ouzou, 20/10/2013.

[32] A. Gherboudj, « Méthodes de résolution de problèmes difficiles académiques », thèse de doctorat, Université de Constantine2, 2013.

[33] http://www.learn.geekinterview.com/database/oracle/advantages-of-using -oracle.html, consultu le : 13/04/2016.

[34] H. E. Seribli, « développement et implémentaion d'un solveur Bio_inspiré pour l'alignement de séquences Biologiques », mémoire de master, université de m'sila ,14/06/2015.

الملخص

ظهرت مستودعات البيانات كحل قادر على تلبية احتياجات تخزين وتحليل أحجام كبيرة من البيانات في حين تجزئة البيانات هي واحدة من التقنيات المستخدمة لتسريع تنفيذ الاستعلام وتسهيل إدارة مستودع البيانات. أفضل طريقة للخروج من مستودع البيانات العلائقية هي تجزئة جداول البعد ثم استخدام أنماط تجزئة لتقسيم جدول حقيقة. فضاء البحث لتحديد نمط تجزئة الأمثل يمكن أن يكون هامة جدا.

في هذه الدراسة المخصصة بمذكرة نهاية الدراسة نحن نحاول حل مشكلة اختيار نمط تجزئة الأمثل مثل الخوارزمية الجينية والبيانات المستخدمة هي بيانات بنك اختبار benchmark.

الكلمات المفتاحية: بنك اختبار benchmark، مستودعات البيانات، الخوارزمية الجينية

Abstract

Data warehouses have emerged as a solution potential to meet the needs of the storage and analysis of large data volumes while Fragmentation of data is one of the techniques used to speed up query execution and facilitate the management of the data warehouse. The best way to break a relational data warehouse involves first breaking down the dimension tables then to use fragmentation patterns to partition the fact table. The search space to select the optimal fragmentation pattern can be very important.

In this final thesis study, we try to solve the problem of selection of an optimal fragmentation pattern by genetic algorithm and the data provided by benchmark test bench.

Keyword: Benchmark test bank, data warehouses, genetic algorithm.

Résumé

Les entrepôts de données ont émergé comme solution potentielle répondant aux besoins du stockage et de l'analyse de grands volumes de données alors que La fragmentation de donnée est une des techniques utilisée pour 'accélérer l'exécution des requêtes et de faciliter la gestion des données de l'entrepôt. La meilleure manière de fragmenter un entrepôt de données relationnel consiste d'abord à décomposer les tables de dimension ensuite à utiliser des schémas de fragmentation pour partitionner la table de faits. L'espace de recherche pour sélectionner le schéma de fragmentation optimal peut être très important.

Dans ce mémoire de fin d'étude on essayer de résolu le problème de sélection d'un schéma de fragmentation optimal par les algorithmes génétique ainsi que les données utilisé fournis par banc essai benchmark.

Mots clé : Banc d'essai benchmark, les entrepôts de données, les algorithmes génétiques.