الجمهورية الشعبية الديموقراطية الجزائرية
République Algérienne Démocratique et Populaire
وزارة العليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي ٠8 ماي 1945٠ بسيدي بلعباس
École Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbès

THESIS

To obtain the diploma of **Master**

Field: **Computer Science**

Specialty: **Système d'Information et Web (SIW)**

Theme

Query optimization techniques in data warehouses: Study and Comparison

---

Presented by:

Boulma Dacine Youssra

Mimouni Wafaa

Submission Date : **July, 2023** In front of the jury composed of

Mr. Kechar Mohammed	Supervisor
Mme. Naoum Hanae	President
Mr. Awad Samir	Examiner

ACKNOWLEDGEMENT

First and foremost, we would like to express our gratitude to Almighty Allah for granting me the courage and patience to complete this work.

We would like to acknowledge our supervisor, Mr. Kechar, for his valuable guidance throughout this journey.

We extend our thanks to my school, the "École Supérieure en Informatique 08 Mai 1945," and our teachers for their kindness and knowledge sharing. Their commitment to education and guidance have been immeasurable.

To all who have contributed directly or indirectly, We appreciate your support, advice, and presence in making this thesis a success.

Thank you all sincerely for your indispensable contributions. It has been an honor to complete this thesis with your support.

-Wafaa & Dacine

DEDICATION

I dedicate this modest work to:

To the woman who suffered without letting me suffer, who spared no effort to make me happy, my adorable mother, thank you for everything.

To the man who helped me become who I am today, my dear dad Fateh, may God keep and protect him.

To my brothers Abdellah, Younes, and specially to my dearest sister Fatoum.

All my family members, aunts, uncles, my lovely grand mother and my friends who give me love, hope and liveliness. A very big thank you.

My colleagues at school Imene, Djihene, Manel, Ines, Khaoula and Hinda Thank you for the excellent moments spent throughout these five years, the atmosphere and the cultural exchanges.
Words are not enough to express all the good I feel! Just Thank you !!!

- Wafaa

DEDICATION

I dedicate this modest work to:

To the memory of my beloved grandmother and father, who are no longer with us,

To my mother, aunt, and brothers Aghiles, Youcef and Juba for their constant presence, encouragement, and support throughout my journey,

To my lovely cousins Alya and Dassyne,

To my dear friends, especially Kenza, who have played a crucial role in my journey. Their unwavering support and genuine friendship have been a blessing. I am grateful for their care and friendship,

Last but certainly not least, to Amine, who has been my unwavering support throughout every challenge and triumph. Your constant presence has provided me with strength. Having you by my side has made all the difference.

-Dacine

ABSTRACT

The query optimizer is a critical component of database systems, where the adoption of cost-based optimizers is prevalent. A cost-based optimizer employs a plan enumeration algorithm to find the most efficient plan by evaluating its cost. In the cost model, cardinality, the number of tuples through an operator, plays a vital role. Inaccurate cardinality estimation, errors in the cost model, and the vast plan space can hinder the optimizer's ability to find the optimal execution plan for complex queries within a reasonable time frame. Several causes behind these limitations push studies to propose techniques to enhance the quality of cardinality estimation, the major key component of the query optimizer and emphasize the need for continued research to address the challenges associated with cardinality estimation and to develop robust techniques that can adapt to different environment and conditions. The objective of this work is to examine different approaches, including synopsis-based methods, sampling-based methods, and learning-based methods. Specifically, both supervised and unsupervised learning methods are investigated. A comparative analysis reveals that while progress has been made in improving cardinality estimation, consistent enhancements are not always achieved.
Keys words : Query optimizer. Cardinality estimation.

RÉSUMÉ

L'optimisateur de requête est une composante critique des systèmes de bases de données où l'adoption d'optimiseurs basés sur le coût est prédominante. Un optimisateur basé sur les coûts utilise un algorithme de numérotation de plan pour trouver le plan le plus efficace en évaluant son coût. Dans le modèle de coût, la cardinalité, le nombre de tuples à travers un opérateur, joue un rôle vital. Des estimations de cardinalité inexactes, des erreurs dans le modèle de coût et l'espace de planification étendu peuvent entraver la capacité de l'optimisateur à trouver le plan d'exécution optimal pour les requêtes complexes dans un délai raisonnable. Plusieurs causes derrière ces limites poussent les études à proposer des techniques visant à améliorer la qualité de l'estimation de la cardinalité, la principale composante clé du optimisateur de requête, et soulignent la nécessité de poursuivre des recherches afin de relever les défis associés à l'évaluation des cardinalités et de développer des techniques robustes qui peuvent s'adapter à différents environnements et conditions. L'objectif de ce travail est d'examiner différentes approches, y compris les méthodes basées sur le résumé, les méthodes basées sur l'échantillonnage et les méthodes basées sur l'apprentissage. Une analyse comparative révèle que si des progrès ont été réalisés dans l'amélioration de l'estimation de la cardinalité, des améliorations cohérentes ne sont pas toujours réalisées.

Mots clés: Optimiseur de requête. Estimation de la cardinalité.

الملخص

مُحسِّن الاستعلامات هو مكون حاسم في أنظمة قواعد البيانات، حيث تعتبر استخدام المُحسِّنات القائمة على التكلفة شائعة في الأنظمة الحالية لقواعد البيانات. يستخدم مُحسِّن الاستعلامات خوارزمية تعداد الخطط للعثور على الخطة الأكثر كفاءة من خلال تقييم تكلفتها. في نموذج التكلفة، تلعب الانتشارية، أي عدد السجلات من خلال المشغل، دورًا حيويًا. تقدير الانتشارية غير الدقيق والأخطاء في نموذج التكلفة وتعدد خيارات التخطيط يمكن أن تعيق قدرة مُحسِّن الاستعلامات على العثور على الخطة التنفيذية الأمثل للاستعلامات المعقدة في إطار زمني معقول. يدفع هذه القيود العديدة لإقتراح تقنيات لتعزيز جودة تقدير الانتشارية، وهي المكون الرئيسي لمُحسِّن الاستعلامات، وتؤكد على ضرورة مواصلة البحث للتغلب على التحديات المرتبطة بتقدير الانتشارية وتطوير تقنيات قوية يمكنها التكيف مع بيئات وظروف مختلفة. يهدف هذا العمل إلى دراسة مقاربات مختلفة، بما في ذلك الأساليب القائمة على الملخصات والأساليب القائمة على العينات والأساليب القائمة على التعلم. على وجه التحديد، يتم استكشاف كل من الأساليب المعلمة وغير المعلمة للتعلم. يكشف التحليل المقارن عن تحقيق تقدم في تحسين تقدير الانتشارية، ولكن ليس دائمًا تحقيق تحسينات مستدامة.

كلمات مفتاحية :محسِّن الاستعلامات. تقدير الانتشارية